Variation in the Perception and Production of Dialect Vowels in Gavaṛ, Armenia

by

Emma Lucy Portugal

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Linguistics)
in the University of Michigan
2026

Doctoral Committee:

Associate Professor Savithry Namboodiripad, Chair
Professor Emerita Patrice Speeter Beddor
Professor Jelena Krivokapić
Assistant Professor Michael Pifer

Emma Lucy Portugal

esantelm@umich.edu

ORCID iD: 0000-0002-3491-8245

© Emma Lucy Portugal 2026

Acknowledgements

This work would not have been possible without the support of numerous individuals and institutions both at the University of Michigan and in Armenia. First, I would like to thank my advisor, Savithry Namboodiripad, for her outstanding mentorship over the last five years and guidance throughout the dissertation writing process. I would also like to thank my dissertation committee members for generously sharing their expertise – Jelena Krivokapić for helping with various aspects of phonetic analysis, Patrice Speeter Beddor for sharing her vast knowledge of speech perception, and Michael Pifer for encouraging me to think about how my work fits into the study of the Armenian language more broadly.

I would like to express my deepest gratitude to my research assistant, Anahit Petrosyan, whose contributions to this project cannot be overstated. Both phases of data collection in Gavaṛ were made possible by Anahit's tireless work – recruiting participants, conducting interviews, and dealing with all sorts of logistical matters – and her transcribing of interviews greatly facilitated the data analysis process. I would also like to thank Armine Petrosyan for assistance with editing the elicitation materials and consent forms used in this project, and for first introducing me to Gavaṛ's dialect back in 2021. I am also grateful to Mariam Hambardzumyan and all the staff at the Geghark^{h}unik^{h} Regional Library in Gavaṛ, who generously welcomed us and provided a space for conducting the 2025 data collection. Of course, I would also like to thank all my research participants for welcoming me into their city and sharing their language and stories.

Financial support for this project was provided by the Center for Armenian Studies at the University of Michigan. I am grateful to the Center both for this support and for fostering such a lively and supportive Armenian Studies community at the university. This project benefitted from feedback from the participants of the Multidisciplinary Workshop for Armenian Studies (some of whom also piloted the perception experiment), as well as from discussions in my advisor's and committee members' lab meetings.

Finally, I would like to thank my husband, Edwin, and our cat, Big Boy, for always being by my side and believing in me.

Table of Contents

Acknowledgements ii

List of Tables x

List of Figures xxiii

List of Appendices xxxvii

Abstract xxxviii

Chapter 1 Introduction 1

1.1 Chapter summary 1

1.2 Research questions and components of dissertation 1

1.3 Background information – sociolinguistic context and vowel system of Gavaṛ 4

1.3.1 Sociolinguistic context – Armenian language and dialectology 4

1.3.2 Bayazet dialect 9

1.4 Overview of motivations from previous literature 15

1.4.1 Influence of social and phonological factors on production and perception 15

1.4.2 Implications for understanding distinction between vowels and concepts such as "merger" 16

1.5 Outline of dissertation structure 18

Chapter 2 Maintenance of Previously Described Dialect Vowels in Production: Comparison of Demographic Patterns in 2022 and 2025 Data 21

2.1 Chapter summary 21

2.2 Goals and research questions of production studies 22

2.3 Summary of 2022 methods and results 23

2.3.1 Data collection 23

2.3.2 Analysis of maintenance of vowel distinctions 24
2.3.3 Results 37
2.4 Methods of 2025 production study 44
2.4.1 Participants 44
2.4.2 Materials and data collection 46
2.4.3 Analysis 49
2.5 Results 56
2.5.1 [a]/[ɒ]A 58
2.5.2 [ɔ]/[ɒ]O 67
2.5.3 [ɒ]A/[ɒ]O 74
2.6 Comparison and discussion of 2022 and 2025 production data 80
2.6.1 Differences in patterns of social variation – sampling issues 81
2.6.2 Potential effects of token sample sizes and recording quality 85
2.6.3 Potential effects of elicitation methods 87
2.7 General summary of production results 91
Chapter 3 Motivations for Perception Experiment and Overview of Methods and Procedures 93
3.1 Chapter summary 93
3.2 Perception experiment research questions 93
3.3 Motivations from 2022 production results 94
3.4 Motivations from previous literature 104
3.4.1 Vowel mergers, near-mergers, and production-perception links 104
3.4.2 Perception and production of coarticulation and allophonic variation 114
3.4.3 The role of social information in speech perception 118
3.4.4 Summary of contributions of the perception experiment 120
3.5 Overview of 2025 study methods and procedures 123

3.5.1 Methods 123

3.5.2 Procedures 124

Chapter 4 AX Discrimination Task: Evidence for Influence of Phonological Characteristics on Discrimination 128

4.1 Chapter summary 128

4.2 Goals and expectations of the AX task 129

4.3 AX task – stimuli, task structure, and data collection 131

4.3.1 Types of stimuli 131

4.3.2 Creation of stimuli 134

4.3.3 Task structure 141

4.3.4 Participants and data collection 148

4.4 Overview of analyses 148

4.4.1 Criteria for exclusion of participants 148

4.4.2 Summary of analyses and predictions 149

4.5 Analysis and results – individual metrics of performance and linear regression models 151

4.5.1 Calculating metrics of individual performance 151

4.5.2 Linear regression models – predicting metrics of performance with demographics and production metrics 156

4.5.3 Results of analysis of individual metrics of performance 159

4.6 Analysis and results – logistic regression models 177

4.6.1 Logistic regression models – predicting responses to various intervals 177

4.6.2 Results of logistic regression models 180

4.7 General discussion 190

4.7.1 Comparison of results for the two vowel pairs 190

4.7.2 Production metrics and demographic factors 193

4.7.3 Methodological considerations 194

Chapter 5 Dialect Categorization Task: Social Salience of Vowels may Facilitate Categorization 196

5.1 Chapter summary 196

5.2 Goals and expectations of the dialect categorization task 196

5.3 Dialect categorization task – stimuli, task structure, and data collection 198

5.3.1 Types of stimuli 198

5.3.2 Creation of stimuli 201

5.3.3 Task structure 205

5.3.4 Participants and data collection 212

5.4 Overview of analyses 212

5.4.1 Criteria for exclusion of participants 212

5.4.2 Summary of analyses and predictions 213

5.5 Analysis and results – Group-level logistic regression models 215

5.5.1 Logistic regression models – predicting responses to different steps 215

5.5.2 Results of logistic regression models 217

5.6 Analysis and results – characterizing and predicting individual-level metrics and response patterns 226

5.6.1 Individual logistic regression models – calculating slopes and perceptual boundary locations 226

5.6.2 Categorizing participants based on response pattern 229

5.6.3 Linear and logistic regression models – predicting slopes, response patterns, and boundary locations with AX performance, production metrics, demographics 235

5.6.4 Results of linear and logistic regression models 239

5.7 General Discussion 257

5.7.1 Group-level logistic regression models 257

5.7.2 Linear and logistic regression models predicting individual metrics 258

5.7.3 Discussion of overall task performance 261

Chapter 6 Comparison of Two Perception Tasks and Comparison of Perception and Production 264

6.1 Chapter summary 264

6.2 Goals and expectations 265

6.2.1 Comparison of perception in AX and dialect categorization tasks 265

6.2.2 Comparison of perception and production 266

6.3 Summary and comparison of results of relevant analyses presented in previous chapters 268

6.3.1 General comparison of the two vowel pairs in results of AX and dialect categorization tasks 268

6.3.2 Findings related to quantitative comparison of the two perception tasks – weak and contradictory effects 271

6.3.3 Comparison of production and perception 273

6.4 Additional qualitative analyses 278

6.4.1 Additional comparison of individual-level performance across the two perception tasks 279

6.4.2 Comparison of perception/production with discriminability/merger criteria and individual-level perception-production comparison 292

6.5 General summary and discussion 303

6.5.1 Comparison of performance on the two perception tasks 303

6.5.2 Comparison of perception and production, and implications for defining merger 305

Chapter 7 Conclusion 312

7.1 Chapter summary 312

7.2 Implications for broad research questions 312

7.2.1 Influence of social and phonological factors on perception and production 312

7.2.2 Defining vowel distinction / implications for concepts of “merger” and “near-merger” 322

7.3 Methodological contributions 325

7.3.1 Acoustic analysis – potential implications of Bark conversion 325

7.3.2 Methodological implications of comparison of production analyses – participant sampling, token sample size, recording quality, elicitation methods 326

7.3.3 Implications of choices in data processing and statistical analyses 328

7.4 Contributions to documentation of contemporary Armenian dialects 329

7.5 Future directions 333

7.6 Conclusion 334

Appendices 336

Bibliography 363

List of Tables

Table 2.1: Results of linear models in relation to GENDER. P-values of less than 0.05 are italicized. All values are rounded to three decimal places. 37

Table 2.2: Results of the linear regression model predicting Euclidean distance between cluster centers in the analysis of [æ] and [a]. The outlier (Speaker 30) has been excluded from this model. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. The adjusted R-squared of the model was 0.268. 44

Table 2.3: Summary of participant information with respect to GENDER and EDUCATION LEVEL..... 45

Table 2.4: Selected words from the word list. Those words used as perception stimuli in either perception task, along with their near-minimal pairs that were used in the creation of the perception stimuli, are bolded. Note that the 2025 word list also contained the word [baɾt^{h}] ('complex') as a more accurate pair for [mɒɾt^{h}]/[maɾt^{h}], but this word could not be used in the creation of perception stimuli, as it was not included in the 2022 word list. 48

Table 2.5: Median, maximum, and minimum sample sizes among speakers by vowel category and in total. Categories are grouped into those that were and were not included in the analysis. 51

Table 2.6: Median, mean, maximum and minimum percentages of tokens in each category that potentially show the expected diphthongization pattern. 56

Table 2.7: Results of the linear regression model predicting Pillai scores for [a]/[ɒ]$_{A}$. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. The adjusted R-squared of the model was 0.102. 59

Table 2.8: Results of the linear regression model predicting ED-median for [a]/[ɒ]$_{A}$ calculated with non-Bark-converted data. The model included PREVIOUS and its interaction with BIRTH YEAR. PREVIOUS was dummy coded (such that the coefficients and p-values for PREVIOUS and BIRTH YEAR do not represent main effects). P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. The adjusted R-squared of the model was 0.089. 62

Table 2.9: Results of the linear regression model predicting ED-median for [a]/[ɒ]$_{A}$ calculated with non-Bark-converted data. The model included PREVIOUS. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. The adjusted R-squared of the model was 0.046. 62

Table 2.10: Results of the linear regression model predicting ED-median for [a]/[ɒ]$_{A}$ calculated with Bark-converted data. The model included PREVIOUS and its interaction with

BIRTH YEAR. PREVIOUS was dummy coded (such that the coefficients and p-values for PREVIOUS and BIRTH YEAR do not represent main effects). P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. The adjusted R-squared of the model was 0.136. 64

Table 2.11: Results of the linear regression model predicting ED-median for [a]/[ɒ]$_A$ calculated with Bark-converted data. The model included PREVIOUS. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. The adjusted R-squared of the model was 0.097. 65

Table 2.12: Results of the linear regression model predicting ED-median for [a]/[ɒ]$_A$ calculated with Bark-converted data. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. The adjusted R-squared of the model was 0.047. 65

Table 2.13: Results of the linear regression model predicting Pillai scores for [ɔ]/[ɒ]$_O$. The model included PREVIOUS and its interaction with GENDER. Both predictors were dummy coded, such that the coefficient/p-value for GENDER represents the difference between non-previous men and women, while the coefficient/p-value for PREVIOUS represents the difference between previous women and non-previous women. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. The adjusted R-squared of the model was 0.073. ... 68

Table 2.14: Results of the linear regression model predicting Pillai scores for [ɔ]/[ɒ]$_O$. The model included PREVIOUS. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. The adjusted R-squared of the model was 0.041. 68

Table 2.15: Results of the linear regression model predicting ED-median for [ɔ]/[ɒ]$_O$. The model included PREVIOUS and its interactions with GENDER and BIRTH YEAR. As the categorical predictors were dummy coded, the individual coefficients/p-values for the predictors involved in the interactions do not represent main effects. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. The adjusted R-squared of the model was 0.140. 71

Table 2.16: Results of the linear regression model predicting ED-median for [ɔ]/[ɒ]$_O$. The model included PREVIOUS. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. The adjusted R-squared of the model was -0.061. 71

Table 2.17: Results of the linear regression model predicting Pillai scores for [ɒ]$_A$/[ɒ]$_O$. The model included the interaction of GENDER and EDUCATION LEVEL. Both predictors were dummy coded, such that the coefficient/p-value for GENDER represents the difference between men and women with higher education, while the coefficient/p-value for EDUCATION LEVEL represents the difference between women with and without higher education. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. The adjusted R-squared of the model was 0.394. 74

Table 2.18: Results of the linear regression model predicting Pillai scores for [ɒ]$_A$/[ɒ]$_O$. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. The adjusted R-squared of the model was 0.367. 75

Table 2.19: Results of the linear regression model predicting ED-median for [ɒ]$_A$/[ɒ]$_O$ calculated with non-Bark-converted data. The model included PREVIOUS. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. The adjusted R-squared of the model was 0.658. 76

Table 2.20: Results of the linear regression model predicting ED-median for [ɒ]$_A$/[ɒ]$_O$ calculated with Bark-converted data. The model included PREVIOUS. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. The adjusted R-squared of the model was 0.596. 77

Table 4.1: Test and training/filler stimuli of the AX task. The "word with desired consonantal context" and "word providing other endpoint" columns represent the actual tokens obtained from the talker's sociolinguistic interview data, while the "resulting continuum" column represents the continua endpoints that were obtained after resynthesis. 136

Table 4.2: Original vowel durations and modified vowel durations in milliseconds for the six test words used in the AX and dialect categorization tasks (the words [marɪ^h]/[mɒrɪ^h] ('person') and [χɔz]/[χɒz] ('pig') were used only in the dialect categorization task). 138

Table 4.3: Maximum formant value setting used in the creation of each continuum for the AX task with Winn's Praat script. 139

Table 4.4: Mean F1 values in Hz and Bark for each step of the [pat]/[pɒt], [mas]/[mɒs], [p^{h}ɔs]/[p^{h}ɒs], and [t͡sɔt͡s^h]/[t͡sɒt͡s^h] continua. 140

Table 4.5: Mean F2 values in Hz and Bark for each step of the [pat]/[pɒt], [mas]/[mɒs], [p^{h}ɔs]/[p^{h}ɒs], and [t͡sɔt͡s^h]/[t͡sɒt͡s^h] continua. 140

Table 4.6: F1 and F2 differences (Bark) between consecutive continuum steps for each test word in the AX task. 140

Table 4.7: Minimal pairs used as control stimuli in the AX task. 141

Table 4.8: Training stimuli for the AX task. Where relevant, the "word" column shows the predicted SEA and dialect pronunciations, which correspond to the two ends of the continuum for the given word. The order of the stimuli in the "pair" column reflects the order in which they were presented. The overall order shown below does not reflect the pseudo-randomized order used in the task. 144

Table 4.9: Summary of structure of the test portion of the AX task. See the text for detailed explanations of where each number comes from. 147

Table 4.10: Results of mixed-effects linear regression model predicting ACCURACY for [a]/[ɒ]$_A$. The model included random intercepts for participants. The model structure was as follows: **ACCURACY ~ WORD + EDUCATION LEVEL + BIRTH YEAR + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. 163

Table 4.11: Results of mixed-effects linear regression model predicting ACCURACY for [ɔ]/[ɒ]O. The model included random intercepts for participants. The model structure was as follows: **ACCURACY ~ WORD + HATSARAT + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. 163

Table 4.12: Results of mixed-effects linear regression model predicting ACCURACY for pooled data with WORD predictor. WORD was dummy coded, with *mas* as the reference level. The model included random intercepts for participants. The structure of the model was as follows: **ACCURACY ~ WORD + EDUCATION LEVEL + BIRTH YEAR + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. This table presents p-values from the *lmer()* output, while Table 4.13 below presents Tukey-adjusted p-values from *emmeans()* pairwise comparisons for WORD. 164

Table 4.13: Results of pairwise comparisons of the levels of WORD from the model presented in **Table 4.12**, computed using *emmeans()*. P-values have been adjusted using the Tukey method. P-values of less than 0.05 are italicized. 164

Table 4.14: Results of mixed-effects linear regression model predicting ACCURACY for pooled data with VOWEL PAIR predictor. The model included random intercepts for participants. The structure of the model was as follows: **ACCURACY ~ VOWEL PAIR + EDUCATION LEVEL + BIRTH YEAR + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. 165

Table 4.15: Results of mixed-effects linear regression model predicting ACCURACY for pooled data with VOWEL PAIR predictor and interactions between VOWEL PAIR, EDUCATION LEVEL, BIRTH YEAR, and HATSARAT. The model included random intercepts for participants. The structure of the model was as follows: **ACCURACY ~ VOWEL PAIR + EDUCATION LEVEL + BIRTH YEAR + HATSARAT + VOWEL PAIR:EDUCATION LEVEL + VOWEL PAIR:BIRTH YEAR + VOWEL PAIR:HATSARAT + (1 | participant)**. Categorical predictors were dummy coded, such that the coefficients for predictors involved in interactions do not represent main effects. The main effect of VOWEL PAIR was tested using *emmeans()*, and the difference between the vowel pairs was not significant (p=0.121). P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. 165

Table 4.16: Results of mixed-effects linear regression model predicting CONSISTENT INTERVALS ACCURACY for [a]/[ɒ]A. The model included random intercepts for participants. The structure of the model was as follows: **CONSISTENT INTERVALS ACCURACY ~ WORD + EDUCATION LEVEL + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. 166

Table 4.17: Results of mixed-effects linear regression model predicting CONSISTENT INTERVALS ACCURACY for [ɔ]/[ɒ]O. The model included random intercepts for participants. The structure of the model was as follows: **CONSISTENT INTERVALS ACCURACY ~ WORD + EDUCATION LEVEL + HATSARAT + BIRTH YEAR + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. 167

Table 4.18: Results of mixed-effects linear regression model predicting CONSISTENT INTERVALS ACCURACY for pooled data with WORD predictor. WORD was dummy coded, with *mas* as the reference level. The model included random intercepts for participants. The structure of the model was as follows: **CONSISTENT INTERVALS ACCURACY ~ WORD + EDUCATION LEVEL + BIRTH YEAR + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. This table presents p-values from the *lmer()* output, while Table 4.19 below presents Tukey-adjusted p-values from *emmeans()* pairwise comparisons for WORD. 167

Table 4.19: Results of pairwise comparisons of the levels of WORD from the model presented in **Table 4.18**, computed using *emmeans()*. P-values have been adjusted using the Tukey method. P-values of less than 0.05 are italicized. 168

Table 4.20: Results of mixed-effects linear regression model predicting CONSISTENT INTERVALS ACCURACY for pooled data with VOWEL PAIR predictor. The model included random intercepts for participants. The structure of the model was as follows: **CONSISTENT INTERVALS ACCURACY ~ VOWEL PAIR + EDUCATION LEVEL + BIRTH YEAR + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. 168

Table 4.21: Results of mixed-effects linear regression model predicting CONSISTENT INTERVALS ACCURACY for pooled data with VOWEL PAIR predictor and interactions between VOWEL PAIR, EDUCATION LEVEL, BIRTH YEAR, and HATSARAT. The model included random intercepts for participants. The structure of the model was as follows: **CONSISTENT INTERVALS ACCURACY ~ VOWEL PAIR + EDUCATION LEVEL + BIRTH YEAR + HATSARAT + VOWEL PAIR:EDUCATION LEVEL + VOWEL PAIR:BIRTH YEAR + VOWEL PAIR:HATSARAT + (1 | participant)**. Categorical predictors were dummy coded, such that the coefficients for predictors involved in interactions do not represent main effects. The main effect of VOWEL PAIR was tested using *emmeans()*, and the difference between the vowel pairs was not found to be significant (p=0.134). P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. 168

Table 4.22: Results of mixed-effects linear regression model predicting *d'* for [a]/[ɒ]A. The model included random intercepts for participants. The structure of the model was as follows: ***d'*** **~ WORD + BIRTH YEAR + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. 169

Table 4.23: Results of mixed-effects linear regression model predicting *d'* for [ɔ]/[ɒ]O. The model included random intercepts for participants. The structure of the model was as follows: ***d'*** **~ WORD + BIRTH YEAR + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. 169

Table 4.24: Results of mixed-effects linear regression model predicting *d'* for pooled data with WORD predictor. WORD was dummy coded, with *mas* as the reference level. The model included random intercepts for participants. The structure of the model was as follows: ***d'*** **~ WORD + BIRTH YEAR + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. This table presents p-values from the *lmer()* output,

while Table 4.25 below presents Tukey-adjusted p-values from *emmeans()* pairwise comparisons for WORD. 170

Table 4.25: Results of pairwise comparisons of WORD from the model presented in **Table 4.24**, computed using *emmeans()*. P-values have been adjusted using the Tukey method. P-values of less than 0.05 are italicized. 170

Table 4.26: Results of mixed-effects linear regression model predicting *d'* for pooled data with VOWEL PAIR predictor. The model included random intercepts for participants. The structure of the model was as follows: ***d'* ~ VOWEL PAIR + BIRTH YEAR + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. 171

Table 4.27: Results of mixed-effects linear regression model predicting SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED in Bark for [a]/[ɒ]$_A$. The model included random intercepts for participants. The structure of the model was as follows: **SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED ~ WORD + EDUCATION LEVEL + BIRTH YEAR + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. 172

Table 4.28: Results of mixed-effects linear regression model predicting SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED in Bark for [ɔ]/[ɒ]$_O$. The model included random intercepts for participants. The structure of the model was as follows: **SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED ~ WORD + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. 172

Table 4.29: Results of mixed-effects linear regression model predicting SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED in Bark for pooled data with WORD predictor. WORD was dummy coded, with *mas* as the reference level. The model included random intercepts for participants. The structure of the model was as follows: **SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED ~ WORD + BIRTH YEAR + EDUCATION LEVEL + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. This table presents p-values from the *lmer()* output, while Table 4.30 below presents Tukey-adjusted p-values from *emmeans()* pairwise comparisons for WORD. 172

Table 4.30: Results of pairwise comparisons of the levels of WORD from the model presented in **Table 4.29**, computed using *emmeans()*. P-values have been adjusted using the Tukey method. P-values of less than 0.05 are italicized. 173

Table 4.31: Results of mixed-effects linear regression model predicting SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED in Bark for pooled data with VOWEL PAIR predictor. The model included random intercepts for participants. The structure of the model was as follows: **SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED ~ VOWEL PAIR + BIRTH YEAR + EDUCATION LEVEL + (1 | participant).** P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. 174

Table 4.32: Results of mixed-effects linear regression model predicting SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED in Bark for pooled data with VOWEL PAIR predictor and interactions between VOWEL PAIR, EDUCATION, and BIRTH YEAR. The model included random intercepts for participants. The structure of the model was as follows: **SMALLEST INTERVAL**

CONSISTENTLY DISCRIMINATED ~ VOWEL PAIR + BIRTH YEAR + EDUCATION LEVEL + VOWEL PAIR:BIRTH YEAR + VOWEL PAIR:EDUCATION LEVEL + (1 | participant). Categorical predictors were dummy coded, such that the coefficients for predictors involved in interactions do not represent main effects. The main effect of VOWEL PAIR was tested using *emmeans()*, and a significant difference was found between the vowel pairs ($p<0.001$). P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. 174

Table 4.33: Summary of results of models predicting ACCURACY. 175

Table 4.34: Summary of results of models predicting CONSISTENT INTERVALS ACCURACY. 175

Table 4.35: Summary of results of models predicting *d'*..... 176

Table 4.36: Summary of results of models predicting SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED in Bark. 176

Table 4.37: Results of mixed-effects logistic regression model predicting responses to consistent intervals for [a]/[ɒ]$_A$. The model structure was as follows: **RESPONSE CORRECT ~ STEP PAIR * WORD + (STEP PAIR * WORD | participant) + (1 | trial order)**. The reference level of WORD was *mas.* P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. 180

Table 4.38: Results of mixed-effects logistic regression model predicting responses to consistent intervals for [ɔ]/[ɒ]$_O$. The model structure was as follows: **RESPONSE CORRECT ~ STEP PAIR * WORD + EDUCATION LEVEL + BIRTH YEAR + (STEP PAIR + WORD | participant) + (1 | trial_order)**. The reference level for WORD was *pos.* P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. 183

Table 4.39: Results of mixed-effects logistic regression model predicting responses to varying intervals for [a]/[ɒ]$_A$. The model structure was as follows: **RESPONSE CORRECT ~ STEP PAIR * WORD + EDUCATION LEVEL + (STEP PAIR * WORD | participant) + (1 | trial_order)**. The reference level for WORD was *mas.* P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. 185

Table 4.40: Results of mixed-effects logistic regression model predicting responses to varying intervals for [ɔ]/[ɒ]$_O$. The model structure was as follows: **RESPONSE CORRECT ~ STEP PAIR * WORD + PREVIOUS + (STEP PAIR * WORD | participant) + (1 | trial_order)**. The reference level for WORD was *pos.* P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. 188

Table 5.1: Maximum formant value setting used in the creation of each continuum with Winn's Praat script. 202

Table 5.2: Mean F1 values in Hz and Bark for each step of the [marɪ^h]/[mɒrɪ^h] and [χɔt]/[χɒt] continua. 202

Table 5.3: Mean F2 values in Hz and Bark for each step of the [marɪ^h]/[mɒrɪ^h] and [χɔt]/[χɒt] continua. 202

Table 5.4: F1 and F2 differences (Bark) between consecutive continuum steps for each word. 203

Table 5.5: Words used as a control condition in the dialect categorization task, in which SEA [h] corresponds with [χ] in Gavaṛ dialect pronunciations. 204

Table 5.6: Lexical items representing words strongly associated with Gavaṛ's dialect and words with a neutral dialect association. 204

Table 5.7: Training stimuli for the dialect categorization task. This table reflects the actual number and order of stimuli trials (no trials were repeated during the training portion). 205

Table 5.8: Summary of the amounts of different types of stimuli in the test portion of the dialect categorization task. 208

Table 5.9: Results of mixed-effects logistic regression model predicting responses to the dialect categorization task for [a]/[ɒ]$_A$. The model structure was as follows: **RESPONSE ~ STEP * WORD + EDUCATION LEVEL + ED-median + (STEP * WORD | participant) + (1 | order within block) + (1 | overall order)**. The reference level for WORD was *mard.* P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. 218

Table 5.10: Results of pairwise comparisons of the levels of WORD from the model presented in **Table 5.9**, computed using *emmeans()*. P-values have been adjusted using the Tukey method. P-values of less than 0.05 are italicized. 218

Table 5.11: Results of pairwise comparisons of the slope of STEP across levels of WORD from the model presented in **Table 5.9**, computed using *emtrends()*. P-values have been adjusted using the Tukey method. P-values of less than 0.05 are italicized. 218

Table 5.12: Slopes of STEP across levels of WORD from the model presented in Table 5.9, computed using *emtrends()*. 0.95 confidence intervals calculated with *emtrends()* are provided. 219

Table 5.13: Results of mixed-effects logistic regression model predicting responses to the dialect categorization task for [ɔ]/[ɒ]$_O$. The model structure was as follows: **RESPONSE ~ STEP * WORD + GENDER + Pillai scores + (STEP * WORD | participant) + (1 | order within block) + (1 | overall order)**. The reference level for WORD was *pos.* P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. 221

Table 5.14: Results of pairwise comparisons of the levels of WORD from the model presented in **Table 5.13**, computed using *emmeans()*. P-values have been adjusted using the Tukey method. P-values of less than 0.05 are italicized. 222

Table 5.15: Results of pairwise comparisons of the slope of STEP across levels of WORD from the model presented in **Table 5.13**, computed using *emtrends()*. P-values have been adjusted using the Tukey method. P-values of less than 0.05 are italicized. 222

Table 5.16: Slopes of STEP across levels of WORD from the model presented in **Table 5.13**, computed using *emtrends()*. 0.95 confidence intervals calculated with *emtrends()* are provided. 223

Table 5.17: Summary of demographic, production, and AX predictors tested in linear and logistic regression models predicting metrics of performance on the dialect categorization task. 236

Table 5.18: Overview of outcome variables of linear and logistic regression models predicting performance on the dialect categorization task. 237

Table 5.19: Results of mixed-effects linear regression model predicting dialect categorization SLOPE for the four stimuli words included in both perception tasks. The reference level of WORD was *mas.* The model included random intercepts for participants. The structure of the model was as follows: **SLOPE ~ WORD + BIRTH YEAR + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. 240

Table 5.20: Results of mixed-effects linear regression model predicting dialect categorization SLOPE for all six stimuli words. The reference level of VOWEL PAIR was [a]/[ɒ]$_A$. The model included random intercepts for participants. The structure of the model was as follows: **SLOPE ~ VOWEL PAIR + BIRTH YEAR + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. 241

Table 5.21: Results of mixed-effects linear regression model predicting dialect categorization SLOPE for all six stimuli words. The model included random intercepts for participants, as well as interactions of VOWEL PAIR with GENDER and BIRTH YEAR. The structure of the model was as follows: **SLOPE ~ VOWEL PAIR + BIRTH YEAR + GENDER + VOWEL PAIR:BIRTH YEAR + VOWEL PAIR:GENDER + (1 | participant)**. The reference level of VOWEL PAIR was [a]/[ɒ]$_A$. Categorical predictors were dummy coded, such that the coefficients for predictors involved in interactions do not represent main effects. The main effect of VOWEL PAIR was tested using *emmeans()*, and a significant difference was found between the vowel pairs (p=0.041). P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. 242

Table 5.22: Results of mixed-effects linear regression model predicting ABSOLUTE VALUE OF SLOPE for the four stimuli words included in both perception tasks. The reference level of WORD was *mas.* The model included random intercepts for participants. The structure of the model was as follows: **ABSOLUTE VALUE OF SLOPE ~ WORD + GENDER + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. 242

Table 5.23: Results of mixed-effects linear regression model predicting ABSOLUTE VALUE OF SLOPE for the four stimuli words included in both perception tasks. The model included random intercepts for participants, as well as interactions between VOWEL PAIR, GENDER and EDUCATION LEVEL. The structure of the model was as follows: **ABSOLUTE VALUE OF SLOPE ~ VOWEL PAIR + GENDER + EDUCATION LEVEL + VOWEL PAIR:EDUCATION LEVEL + VOWEL PAIR:GENDER + (1 | participant)**. The reference level of VOWEL PAIR was [a]/[ɒ]$_A$. Categorical predictors were dummy coded, such that the coefficients for predictors involved in interactions

do not represent main effects. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. 243

Table 5.24: Results of mixed-effects linear regression model predicting ABSOLUTE VALUE OF SLOPE for all six stimuli words. The reference level of VOWEL PAIR was [a]/[ɒ]A. The model included random intercepts for participants. The structure of the model was as follows: **ABSOLUTE VALUE OF SLOPE ~ VOWEL PAIR + GENDER + BIRTH YEAR + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. 244

Table 5.25: Results of mixed-effects linear regression model predicting ABSOLUTE VALUE OF SLOPE for all six stimuli words. The model included random intercepts for participants, as well as interactions of VOWEL PAIR with BIRTH YEAR, EDUCATION LEVEL, and *d'*. The structure of the model was as follows: **ABSOLUTE VALUE OF SLOPE ~ VOWEL PAIR + d' + GENDER + BIRTH YEAR + EDUCATION LEVEL + VOWEL PAIR:d' + VOWEL PAIR:BIRTH YEAR + VOWEL_PAIR:EDUCATION LEVEL + (1 | participant)**. The reference level of VOWEL PAIR was [a]/[ɒ]A. Categorical predictors were dummy coded, such that the coefficients for predictors involved in interactions do not represent main effects. The main effect of VOWEL PAIR was tested using *emmeans()*, and a significant difference was found between the vowel pairs ($p<0.001$). P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. 244

Table 5.26: Results of mixed-effects linear regression model predicting SLOPE among participants with the expected response pattern for the four stimuli words included in both perception tasks. The reference level of WORD was *mas.* The model included random intercepts for participants. The structure of the model was as follows: **SLOPE ~ WORD + GENDER + PREVIOUS + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. 245

Table 5.27: Results of mixed-effects linear regression model predicting SLOPE among participants with the expected response pattern for all six stimuli words. The reference level of VOWEL PAIR was [a]/[ɒ]A. The model included random intercepts for participants. The structure of the model was as follows: **SLOPE ~ VOWEL PAIR + GENDER + PREVIOUS + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. 246

Table 5.28: Results of mixed-effects logistic regression model predicting RESPONSE PATTERN for the four stimuli words included in both perception tasks. The reference level of WORD was *mas.* The model included random intercepts for participants. The structure of the model was as follows: **RESPONSE PATTERN ~ WORD + BIRTH YEAR + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. 247

Table 5.29: Results of mixed-effects logistic regression model predicting RESPONSE PATTERN for the four stimuli words included in both perception tasks. The model included random intercepts for participants, as well as the interaction of VOWEL PAIR and BIRTH YEAR. The structure of the model was as follows: **RESPONSE PATTERN ~ VOWEL PAIR * BIRTH YEAR + (1 | participant)**. The reference level of VOWEL PAIR was [a]/[ɒ]A. Categorical predictors were dummy coded, such that the coefficients for predictors involved in interactions do not represent

main effects. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. 247

Table 5.30: Results of mixed-effects logistic regression model predicting RESPONSE PATTERN for all six stimuli words. The reference level of VOWEL PAIR was [a]/[ɒ]$_A$. The model included random intercepts for participants. The structure of the model was as follows: **RESPONSE PATTERN ~ VOWEL PAIR + BIRTH YEAR + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. 248

Table 5.31: Results of mixed-effects logistic regression model predicting RESPONSE PATTERN for all six stimuli words. The model included random intercepts for participants, as well as interactions of VOWEL PAIR with BIRTH YEAR, GENDER, and AX ACCURACY. The structure of the model was as follows: **RESPONSE PATTERN ~ VOWEL PAIR + GENDER + BIRTH YEAR + ACCURACY + VOWEL PAIR:GENDER + VOWEL PAIR:BIRTH YEAR + VOWEL PAIR:ACCURACY + (1 | participant)**. The reference level of VOWEL PAIR was [a]/[ɒ]$_A$. Categorical predictors were dummy coded, such that the coefficients for predictors involved in interactions do not represent main effects. The main effect of VOWEL PAIR was tested using *emmeans()*, and a significant difference was found between the vowel pairs (p=0.028). P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. 249

Table 5.32: Slopes of AX ACCURACY across vowel pairs from the model presented in **Table 5.31**, computed using *emtrends()*. 0.95 confidence intervals calculated with *emtrends()* are provided. 249

Table 5.33: Results of mixed-effects logistic regression model predicting RESPONSE PATTERN for all six stimuli words. The model included random intercepts for participants, as well as interactions of VOWEL PAIR with BIRTH YEAR, GENDER, and *d'*. The structure of the model was as follows: **RESPONSE PATTERN ~ VOWEL PAIR + GENDER + BIRTH YEAR + d' + VOWEL PAIR:GENDER + VOWEL PAIR:BIRTH YEAR + VOWEL PAIR:d' + (1 | participant)**. The reference level of VOWEL PAIR was [a]/[ɒ]$_A$. Categorical predictors were dummy coded, such that the coefficients for predictors involved in interactions do not represent main effects. The main effect of VOWEL PAIR was tested using *emmeans()*, and a significant difference was found between the vowel pairs (p=0.049). P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. 249

Table 5.34: Slopes of *d'* across vowel pairs from the model presented in **Table 5.33**, computed using *emtrends()*. 0.95 confidence intervals calculated with *emtrends()* are provided. 250

Table 5.35: Results of mixed-effects linear regression model predicting BOUNDARY LOCATION for the four stimuli words included in both perception tasks. The reference level of WORD was *mas.* The model included random intercepts for participants. The structure of the model was as follows: **BOUNDARY LOCATION ~ WORD + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. 251

Table 5.36: Results of mixed-effects linear regression model predicting BOUNDARY LOCATION for the four stimuli words included in both perception tasks. The model included random intercepts for participants, as well as the interactions of VOWEL PAIR with BIRTH YEAR and ED-

median (from non-Bark-converted data). The structure of the model was as follows: **BOUNDARY LOCATION ~ VOWEL PAIR + ED-median + BIRTH YEAR + VOWEL PAIR:BIRTH YEAR + VOWEL PAIR:ED-median + (1 | participant)**. The reference level of VOWEL PAIR was [a]/[ɒ]A. Categorical predictors were dummy coded, such that the coefficients for predictors involved in interactions do not represent main effects. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. 251

Table 5.37: Results of mixed-effects linear regression model predicting BOUNDARY LOCATION for all six stimuli words. The reference level of VOWEL PAIR was [a]/[ɒ]A. The model included random intercepts for participants. The structure of the model was as follows: **BOUNDARY LOCATION ~ VOWEL PAIR + EDUCATION LEVEL + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. 253

Table 5.38: Results of mixed-effects linear regression model predicting BOUNDARY LOCATION for all six stimuli words. The model included random intercepts for participants, as well as interactions of VOWEL PAIR with BIRTH YEAR, EDUCATION LEVEL, and ED-median (calculated from non-Bark-converted data). The structure of the model was as follows: **BOUNDARY LOCATION ~ VOWEL PAIR + BIRTH YEAR + EDUCATION LEVEL + ED-median + VOWEL PAIR:BIRTH YEAR + VOWEL PAIR:EDUCATION LEVEL + VOWEL PAIR:ED-median + (1 | participant)**. The reference level of VOWEL PAIR was [a]/[ɒ]A. Categorical predictors were dummy coded, such that the coefficients for predictors involved in interactions do not represent main effects. The main effect of VOWEL PAIR was tested using *emmeans()*, and a nearly significant difference was found between the vowel pairs (p=0.051). P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. 253

Table 5.39: Summary of key results of models predicting SLOPE. 255

Table 5.40: Summary of key results of models predicting ABSOLUTE VALUE OF SLOPE. 255

Table 5.41: Summary of key results of models predicting SLOPE only among participants with expected response pattern. 256

Table 5.42: Summary of key results of models predicting RESPONSE PATTERN. 256

Table 5.43: Summary of key results of models predicting BOUNDARY LOCATION. 257

Table 6.1: Comparison of participants' produced vowel differences for [a]/[ɒ]A and [ɔ]/[ɒ]O with the three Bark thresholds described in Section 6.4.2.2. The first two rows show the number of participants whose produced F1 and F2 differences were larger than the 0.5 and 0.2 Bark per-formant thresholds (i.e., were discriminable/unmerged). The third row shows the number of participants for whom the Euclidean distance between the relevant vowels in production was larger than the 0.65 Bark threshold (i.e., were unmerged). The fourth row shows the number of participants whose productions would have been deemed indiscriminable/merged by each method. 299

Table 6.2: Comparison of participants' SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED for each word with the three Bark thresholds described in Section 6.4.2.2. The first two rows show the number of participants whose SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED was

smaller in F1 or F2 than the 0.5 and 0.2 Bark per-formant thresholds (F2 is not considered for *pos*/*tsots*, for the reasons described above). The third row shows the number of participants for whom the SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED (Euclidean distance) was smaller than the 0.65 Bark threshold. 300

Table 6.3: The first row shows the number of participants who likely would have discriminated their own produced differences for each vowel pair, if the smallest perceived difference is defined as the SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED in the varying intervals portion of the AX task. The second row shows the number of participants whose produced differences for each vowel pair likely would have been discriminated by anyone in the sample, according to the same analysis method. 301

List of Figures

Figure 1.1: Vowel systems of SEA (left) (see Dum-Tragut 2009; Khach^{h}atryan 1988; Seyfarth et al. 2023) and the Gavaṛ (Bayazet) dialect (right) (based on descriptions of Katvalyan [2016]). Vowels not present in SEA are shown in purple. Not all illustrated distinctions are contrastive. The phonemic status of the schwa is debated (see Allen 1950; Dum-Tragut 2009; Khach^{h}atryan 1988; Khachatrian 1996; Vaux 1998). Some sources (e.g., Dum-Tragut 2009), transcribe the SEA mid vowels as [ɛ] and [ɔ], while Seyfarth et al. (2023: 19) describe the mid-front vowel as varying between [ɛ] and [e] and the mid-back vowel as acoustically closer to [o]. The transcription [ɔ] has been used for the sake of consistency with Portugal (2025). 15

Figure 2.1: Pairs of vowels examined in the 2022 analysis. The vowel system of SEA is shown in blue (see Dum-Tragut 2009; Khach^{h}atryan 1988; Seyfarth et al. 2023), while the additional vowels of Gavaṛ's dialect are shown in purple. 26

Figure 2.2: The [æ] and [a] tokens of a participant exhibiting a merger by transfer of [æ] and [a]. The tokens where [a] was predicted are represented by red dots and a red ellipse (representing 2 standard deviations of this token category), while the blue and black dots represent the tokens where [æ] was predicted. The black dots represent those [æ] tokens that were clustered together with the [a] tokens by k-means analysis. The 3-dimensional density plot represents the distribution of only those tokens where [æ] was predicted. 34

Figure 2.3: The [æ] and [a] tokens of a participant exhibiting a merger by approximation of [æ] and [a]. The tokens where [a] was predicted are represented by red dots and a red ellipse (representing 2 standard deviations of this token category), while the blue and black dots represent the tokens where [æ] was predicted. The black dots represent those [æ] tokens that were clustered together with the [a] tokens by k-means analysis. The 3-dimensional density plot represents the distribution of only those tokens where [æ] was predicted. 35

Figure 2.4: The [æ] and [a] tokens of a participant exhibiting a complete merger of [æ] and [a]. The tokens where [a] was predicted are represented by red dots and a red ellipse (representing 2 standard deviations of this token category), while the blue and black dots represent the tokens where [æ] was predicted. The 3-dimensional density plot represents the distribution of only those tokens where [æ] was predicted. 35

Figure 2.5: Individual median formant values grouped by GENDER in normalized units (without Bark conversion). Confidence ellipses created in *ggplot2* with the *stat_ellipse()* function represent approximately one standard deviation. 38

Figure 2.6: Individual Euclidean distance measurements between the centers of clusters identified by k-means analysis of the tokens where [æ] and [a] were predicted. Measurements are presented in normalized units and color-coded by GENDER. 43

Figure 2.7: Individual Euclidean distance measurements between the centers of clusters identified by k-means analysis of the tokens where [æ] and [a] were predicted. Measurements are presented in normalized units and color-coded by EDUCATION LEVEL. 43

Figure 2.8: Density of participants by birth year. Each dot represents one participant. 46

Figure 2.9: A token of [ɒ]ᴏ with a formant trajectory that could represent the expected pattern of diphthongization (moving from more [ɔ]-like to more [a]-like). The axes are shown in Hz. The three measurements represent the 20-40%, 40-60%, and 60-80% duration intervals of the vowel token. 56

Figure 2.10: Individual median formant values from 2025 grouped by GENDER in normalized units (without Bark conversion). Confidence ellipses created in *ggplot2* with the *stat_ellipse()* function represent approximately one standard deviation. 57

Figure 2.11: Individual median formant values from 2025 grouped by GENDER in normalized units (with Bark conversion). Confidence ellipses created in *ggplot2* with the *stat_ellipse()* function represent approximately one standard deviation. 58

Figure 2.12: Individual Pillai scores by GENDER for **[a]/[ɒ]ᴀ** with superimposed merger/distinction thresholds. 59

Figure 2.13: Individual Pillai scores for **[a]/[ɒ]ᴀ** with superimposed merger/distinction thresholds, coded in accordance with the PREVIOUS predictor. 60

Figure 2.14: Individual vowel plot for Participant 11 (2025). Ellipses represent two standard deviations. 61

Figure 2.15: Individual vowel plot for Participant 43 (2025). Ellipses represent two standard deviations. 61

Figure 2.16: Individual ED-median values for **[a]/[ɒ]ᴀ** calculated with non-Bark-converted data and coded by PREVIOUS. 63

Figure 2.17: Individual ED-median values for **[a]/[ɒ]ᴀ** calculated with non-Bark-converted data and coded by BIRTH YEAR and PREVIOUS, with superimposed regression lines. Each dot represents one participant, the x-axis represents BIRTH YEAR, and the color coding represents PREVIOUS. 64

Figure 2.18: Individual ED-median values for **[a]/[ɒ]ᴀ** calculated with non-Bark-converted data and coded by GENDER. 64

Figure 2.19: Individual ED-median values for **[a]/[ɒ]ᴀ** calculated with Bark-converted data and coded by PREVIOUS. 66

Figure 2.20: Individual ED-median values for **[a]/[ɒ]A** calculated with Bark-converted data and coded by BIRTH YEAR and PREVIOUS, with superimposed regression lines. Each dot represents one participant, the x-axis represents BIRTH YEAR, and the color coding represents PREVIOUS. 67

Figure 2.21: Individual ED-median values for **[a]/[ɒ]A** calculated with Bark-converted data and coded by GENDER. 67

Figure 2.22: Individual Pillai scores for **[ɔ]/[ɒ]O** with superimposed merger/distinction thresholds, coded in accordance with the PREVIOUS predictor. 69

Figure 2.23: Individual Pillai scores by GENDER for **[ɔ]/[ɒ]O** with superimposed merger/distinction thresholds. 70

Figure 2.24: Pillai scores of new (left) and previous (right) participants for the comparison of [ɔ]/[ɒ]O in the 2025 data set. A gender pattern can be seen among the previous participants, but not among the new participants. 70

Figure 2.25: ED-median measurements of new (left) and previous (right) participants for the comparison of **[ɔ]/[ɒ]O** in the 2025 data set. A gender pattern can be seen among the previous participants, but not among the new participants. 72

Figure 2.26: Individual ED-median values for **[ɔ]/[ɒ]O** coded by BIRTH YEAR and PREVIOUS, with superimposed regression lines. Each dot represents one participant, the x-axis represents BIRTH YEAR, and the color coding represents PREVIOUS. 72

Figure 2.27: Individual ED-median values by GENDER for **[ɔ]/[ɒ]O**. 73

Figure 2.28: Individual ED-median values for **[ɔ]/[ɒ]O** coded in accordance with PREVIOUS. ... 73

Figure 2.29: Individual ED-median values by BIRTH YEAR for **[ɔ]/[ɒ]O** with superimposed regression line. 74

Figure 2.30: Individual Pillai scores by GENDER for **[ɒ]A/[ɒ]O** with superimposed merger/distinction thresholds. 75

Figure 2.31: Individual Pillai scores for **[ɒ]A/[ɒ]O** grouped by GENDER and EDUCATION LEVEL with superimposed merger/distinction thresholds. 76

Figure 2.32: Individual ED-median values for **[ɒ]A/[ɒ]O** calculated with non-Bark-converted data and coded by GENDER. 78

Figure 2.33: Individual ED-median values for **[ɒ]A/[ɒ]O** calculated with non-Bark-converted data and coded by PREVIOUS. 79

Figure 2.34: Individual ED-median values for **[ɒ]A/[ɒ]O** calculated with Bark-converted data and coded by GENDER. 79

Figure 2.35: Individual ED-median values for **[ɒ]A/[ɒ]O** calculated with Bark-converted data and coded by PREVIOUS. 79

Figure 2.36: Pillai scores of previous (repeat) participants for [ɔ]/[ɒ]O in the 2025 (left) and 2022 (right) data sets. A similar gender pattern can be seen among these participants across both data sets. Pillai scores are smaller overall in the 2025 data set, likely due to larger sample sizes. Accordingly, although the overlaid merger/distinction thresholds calculated with the method of Stanley and Sneller (2023) differ across the two graphs, a similar number of participants fall above and below the thresholds in both data sets. 82

Figure 2.37: ED-median measurements of previous (repeat) participants for the comparison of [ɔ]/[ɒ]O in the 2025 (left) and 2022 (right) data sets. A similar gender pattern can be seen among these participants across both data sets. 82

Figure 2.38: Pillai scores of previous (repeat) participants for the comparison of [a]/[ɒ]A in the 2025 (left) and 2022 (right) data sets, with merger/distinction thresholds calculated using the method of Stanley & Sneller (2023). 87

Figure 2.39: Individual formant measurements in normalized units for the group of tokens where [æ] was expected (word list only – individual tokens represented by red dots) and the group of tokens where [a] was expected (black ellipses representing two standard deviations from the mean) in 2022. 90

Figure 2.40: Individual formant measurements in normalized units for the group of tokens where [æ] was expected (individual tokens represented by red dots) and the group of tokens where [a] was expected (black ellipses representing two standard deviations from the mean) in 2025 91

Figure 3.1: Individual Pillai scores (2022 data) by GENDER for **[ɔ]/[ɒ]O** with superimposed distinction thresholds. 96

Figure 3.2: Individual Pillai scores (2022 data) by GENDER for **[a]/[ɒ]A** with superimposed distinction thresholds. 96

Figure 3.3: Individual Pillai scores (2022 data) by GENDER for **[ɒ]A/[ɒ]O** with superimposed distinction thresholds. 97

Figure 3.4: Individual vowel plot of Speaker 3 from 2022. 97

Figure 3.5: Individual vowel plot of Speaker 4 from 2022. 98

Figure 3.6: Individual vowel plot of Speaker 11 from 2022. 98

Figure 3.7: Screen shown at the end of the experiment, prompting participants to answer whether or not they thought they knew the person whose voice was used for the stimuli. The screen reads, “Do you think that you know the person whose voice you heard during the study? If yes, press 1. If no, press 0.” 127

Figure 4.1: Instructions displayed at the beginning of the training and test portions of the AX task. The text reads, "You will hear two forms of pronunciation of some words and will try to understand if the ways that those words sound are the same or different. If you think they are the same, press 1. If you think they are different, press 0. Press the spacebar to continue."..... 143

Figure 4.2: Response screen from the AX task. The left option reads "same (1)", while the right option reads "different (0)". 143

Figure 4.3: Participant 2's responses for varying-size intervals the word *pat*. The left and right-hand columns indicate incorrect and correct responses respectively. The interval 4-7 was chosen because the participant reached 75% accuracy at this interval and had 100% accuracy for all the larger intervals. 155

Figure 4.4: Participant 42's responses for varying-size intervals for the word *pat*. The left and right-hand columns indicate incorrect and correct responses respectively. The interval 4-7 was chosen because the participant reached 100% accuracy at that interval (after only having 25% accuracy at the previous interval). Though the participant's accuracy dropped at the subsequent interval 3-7, this did not affect the interval choice because the participant's accuracy at this interval was still 75%. 156

Figure 4.5: Participant 34's responses for varying-size intervals for the word *pat*. The left and right-hand columns indicate incorrect and correct responses respectively. This participant failed to reach consistent 75% accuracy at any interval – although they reached 75% accuracy at the interval 4-7, they dropped to 50% accuracy at the larger intervals 3-7 and 1-7, meaning that their interval was recorded as 1-7+. 156

Figure 4.6: The left panel shows mean ACCURACY for all participants by WORD. Error bars represent ± one standard deviation. The right panel shows violin plots of participants' ACCURACY by word with overlaid boxplots. The box represents the range between the first and third quartiles (interquartile range), while the horizontal line represents the median. The whiskers represent data points that lie up to 1.5 times the interquartile range from the edge of the box, while black dots represent outliers. 160

Figure 4.7: The left panel shows mean CONSISTENT INTERVALS ACCURACY by WORD. Error bars represent ± one standard deviation. The right panel shows violin plots of participants' CONSISTENT INTERVALS ACCURACY by word with overlaid boxplots. The box represents the range between the first and third quartiles (interquartile range), while the horizontal line represents the median. The whiskers represent data points that lie up to 1.5 times the interquartile range from the edge of the box. 160

Figure 4.8: The left panel shows mean *d'* values among participants by WORD. Error bars represent ± one standard deviation. The right panel shows violin plots of participants' *d'* values per word with overlaid boxplots. The box represents the range between the first and third quartiles (interquartile range), while the horizontal line represents the median. The whiskers represent data points that lie up to 1.5 times the interquartile range from the edge of the box, while black dots represent outliers. 161

Figure 4.9: The left panel shows violin plots with overlaid boxplots showing each participant's SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED by WORD. A y-axis value of 7 corresponds to an interval of 1-7+ (i.e., not reaching 75% accuracy at any interval), while a y-axis value of 2 corresponds to the interval 5-7 (no participant achieved consistent 75% accuracy at the interval 6-7 for any word). The box represents the range between the first and third quartiles (interquartile range), while the horizontal line represents the median. The whiskers represent data points that lie up to 1.5 times the interquartile range from the edge of the box. The right panel shows the same data as discrete counts of participants assigned to each numeric interval. 162

Figure 4.10: Violin plots with overlaid boxplots showing each participant's SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED by WORD in Bark. The box represents the range between the first and third quartiles (interquartile range), while the horizontal line represents the median. The whiskers represent data points that lie up to 1.5 times the interquartile range from the edge of the box, while black dots represent outliers. 162

Figure 4.11: Pooled responses by WORD for consistent intervals (2-step intervals) for the [a]/[ɒ]$_A$ vowel continuum. Dark blue represents correct responses (i.e., correctly discriminating the two members of the pair), while green represents incorrect responses (i.e., failing to discriminate the two members of the pair). 181

Figure 4.12: Violin plots, boxplots, and jittered scatter plots of individual responses by WORD for consistent intervals (2-step intervals) for the [a]/[ɒ]$_A$ vowel continuum. Each dot represents one participant's proportion of correct responses for the given step pair for the given word (the points have been jittered both vertically and horizontally so that all points would be visible, which results in some points appearing to be below 0 or above 1). The violin plots and boxplots summarize the distribution of the data shown in the scatter plots. The box represents the range between the first and third quartiles (interquartile range), while the horizontal line represents the median. The whiskers represent data points that lie up to 1.5 times the interquartile range from the edge of the box. 182

Figure 4.13: Pooled responses by WORD for consistent intervals (3-step intervals) for the [ɔ]/[ɒ]$_O$ vowel continuum. Green represents correct responses (i.e., correctly discriminating the two members of the pair), while red represents incorrect responses (i.e., failing to discriminate the two members of the pair). 184

Figure 4.14: Violin plots, boxplots, and jittered scatter plots of individual responses by WORD for consistent intervals (3-step intervals) for the [ɔ]/[ɒ]$_O$ vowel continuum. Each dot represents one participant's proportion of correct responses for the given step pair for the given word (the points have been jittered both vertically and horizontally so that all points would be visible, which results in some points appearing to be below 0 or above 1). The violin plots and boxplots summarize the distribution of the data shown in the scatter plots. The box represents the range between the first and third quartiles (interquartile range), while the horizontal line represents the median. The whiskers represent data points that lie up to 1.5 times the interquartile range from the edge of the box. 184

Figure 4.15: Pooled responses by WORD for varying intervals for the [a]/[ɒ]$_A$ vowel continuum. The right-hand column represents correct responses, while the left-hand column represents incorrect responses (i.e., failure to discriminate the interval). 185

Figure 4.16: Pooled responses by WORD for varying intervals for the [a]/[ɒ]$_A$ vowel continuum. Dark blue represents correct responses, while green represents incorrect responses (i.e., failure to discriminate the interval). 186

Figure 4.17: Proportion of correct responses by EDUCATION LEVEL for varying intervals for the [a]/[ɒ]$_A$ vowel continuum. Purple represents correct responses, while yellow represents incorrect responses (i.e., failure to discriminate the interval). 186

Figure 4.18: Pooled responses by WORD for varying intervals for the [ɔ]/[ɒ]$_O$ vowel continuum. The right-hand column represents correct responses, while the left-hand column represents incorrect responses (i.e., failure to discriminate the interval). 188

Figure 4.19: Pooled responses by WORD for varying intervals for the [ɔ]/[ɒ]$_O$ vowel continuum. Dark blue represents correct responses, while green represents incorrect responses (i.e., failure to discriminate the interval). Recall that the interval 1-7 had extra repetitions for this vowel pair, as explained in Section 4.3.3.3. 189

Figure 5.1: Template of the blocks of the dialect categorization task. Pink represents [h]/[χ] stimuli, orange represents test stimuli, green represents lexical stimuli, and yellow represents [a]/[æ] stimuli. Blue represents cases in which some blocks had lexical stimuli and some blocks had [h]/[χ] stimuli. 209

Figure 5.2: Blocks of the dialect categorization task (each column is one block). Pink represents [h]/[χ] stimuli, orange represents test stimuli, green represents lexical stimuli, and yellow represents [a]/[æ] stimuli. Note that the order of stimuli for the word [t^hæt^h]/[t^hat^h] is opposite from the rest of the words representing the [a]/[æ] continuum because the SEA pronunciation [t^hat^h] (rather than [t^hæt^h]) served as the basis for this continuum (see Section 4.3.2.1), such that the step numbers for this word are opposite from the other [a]/[æ] words. .. 210

Figure 5.3: Instructions displayed at the beginning of the training and test portions of the dialect categorization task. The text reads, “Now you will hear some words one at a time. Considering the given word’s form, pronunciation, and the way it sounds, you must decide if that word’s form / pronunciation / the way it sounds is characteristic of Gavaṛ’s dialect or not. If you think that it is characteristic of Gavaṛ’s dialect, press 1. If you think that it is not characteristic of Gavaṛ’s dialect (that is, not only people from Gavaṛ, but also other residents of Armenia can speak that way), press 0. The same person will pronounce all of the words. So, you should make your decision not based on the person’s voice, but rather based on the word’s form / pronunciation / the way it sounds. Press the spacebar to continue.” 212

Figure 5.4: Response screen from the dialect categorization task. Image from MediaHub (2024). 212

Figure 5.5: Pooled responses by WORD for [a]/[ɒ]$_A$ in the dialect categorization task. Purple (1) represents “Gavaṛ” responses, while yellow (0) represents “not Gavaṛ” responses. 219

Figure 5.6: Proportion of "Gavaṛ" responses by EDUCATION LEVEL for [a]/[ɒ]$_A$ in the dialect categorization task. Purple represents "Gavaṛ" (1) responses, while yellow represents "not Gavaṛ" (0) responses. "Gavaṛ" responses made up a larger portion of the responses for the non-higher group. 220

Figure 5.7: Proportion of "Gavaṛ" responses in relation to ED-median for [a]/[ɒ]$_A$ in the dialect categorization task. Purple represents "Gavaṛ" (1) responses, while yellow represents "not Gavaṛ" (0) responses. Each bar represents one participant, and the bars are arranged from smallest ED-median in production for the given vowel pair (left) to largest ED-median (right). Participant numbers are displayed at the top of the graph. There is a weak visual trend according to which the proportion of "Gavaṛ" responses decreases as ED-median increases. .. 220

Figure 5.8: Pooled responses by WORD for [ɔ]/[ɒ]$_O$ in the dialect categorization task. Purple (1) represents "Gavaṛ" responses, while yellow (0) represents "not Gavaṛ" responses. 223

Figure 5.9: Proportion of "Gavaṛ" responses by GENDER for [ɔ]/[ɒ]$_O$ in the dialect categorization task. Purple represents "Gavaṛ" (1) responses, while yellow represents "not Gavaṛ" (0) responses. "Gavaṛ" responses made up a slightly larger portion of the responses for women in comparison with men. 224

Figure 5.10: Proportion of "Gavaṛ" responses in relation to Pillai scores for [ɔ]/[ɒ]$_O$ in the dialect categorization task. Purple represents "Gavaṛ" (1) responses, while red represents "not Gavaṛ" (0) responses. Each bar represents one participant, and the bars are arranged from smallest Pillai score in production for the given vowel pair (left) to largest Pillai score (right). Participant numbers are displayed at the top of the graph. According to the model presented in Table 5.13, there is a weak trend according to which proportion of "Gavaṛ" responses increases as Pillai scores increases (i.e., as overlap between vowel categories in production decreases). 224

Figure 5.11: Example of shallow response pattern (left – Participant 9, *mard*) vs. "complete separation" response pattern (right – Participant 9, *mas*). Purple (1) represents "Gavaṛ" responses, while yellow (0) represents "not Gavaṛ" responses. Step 1 represents the most dialect-like variant, while step 7 represents the most SEA-like variant. 229

Figure 5.12: Participant 26's responses for *pat.* Purple (1) represents "Gavaṛ" responses, while yellow (0) represents "not Gavaṛ" responses. Step 1 represents the most dialect-like variant, while step 7 represents the most SEA-like variant. Though the computed slope was -0.65, the participant was deemed to have neither the expected nor opposite response pattern. 230

Figure 5.13: Participant 14's responses for *mard.* Purple (1) represents "Gavaṛ" responses, while yellow (0) represents "not Gavaṛ" responses. Step 1 represents the most dialect-like variant, while step 7 represents the most SEA-like variant. Though the computed slope was -0.24, the participant was deemed to have neither the expected nor opposite response pattern due to the fact that the pattern appears random, and the response to the two endpoints of the continuum was equal. 230

Figure 5.14: Participant 2's responses for *tsots.* Purple (1) represents "Gavaṛ" responses, while yellow (0) represents "not Gavaṛ" responses. Step 1 represents the most dialect-like variant, while step 7 represents the most SEA-like variant. Though the computed slope was 0.11, and

the participant's response pattern is shallow and noisy, the participant was deemed to have the "opposite" response pattern, due to the fact that "Gavaṛ" responses were generally higher toward one end of the continuum than the other (in particular, due to the cluster of two low responses at the left end of the continuum). 231

Figure 5.15: All participants' response patterns by word. Green represents the expected pattern, blue represents the opposite pattern, and red represents no pattern (Participant 16 is the participant whose responses to the experiment were lost). The left three columns are words representing [a]/[ɒ]$_A$, while the right three columns represent [ɔ]/[ɒ]$_O$. 232

Figure 5.16: Participants' response patterns by word, arranged by the number of words with the expected pattern and the vowel pair that had more expected response patterns for each participant. The group of participants who had the expected pattern for three words is divided into those whose three expected patterns all belonged to the same vowel pair, and those with more mixed response patterns. 233

Figure 5.17: Violin plots of participants' individual values of SLOPE by WORD with overlaid boxplots. These plots include those participants who were included in the models predicting slope (see Section 5.6.3). More negative slopes correspond to responding in a more expected manner. The box represents the range between the first and third quartiles (interquartile range), while the horizontal line represents the median. The whiskers represent data points that lie up to 1.5 times the interquartile range from the edge of the box, while black dots represent outliers. The left three words represent [a]/[ɒ]$_A$, while the right three words represent [ɔ]/[ɒ]$_O$. 241

Figure 5.18: Violin plots of ABSOLUTE VALUE OF SLOPE by WORD with overlaid boxplots. These plots include those participants who were included in the models predicting slope (see Section 5.6.3). The box represents the range between the first and third quartiles (interquartile range), while the horizontal line represents the median. The whiskers represent data points that lie up to 1.5 times the interquartile range from the edge of the box, while black dots represent outliers. The left three words represent [a]/[ɒ]$_A$, while the right three words represent [ɔ]/[ɒ]$_O$. 245

Figure 5.19: The left panel presents violin plots showing the distribution of SLOPE by WORD among only those participants who had the expected response pattern. The box represents the range between the first and third quartiles (interquartile range), while the horizontal line represents the median. The whiskers represent data points that lie up to 1.5 times the interquartile range from the edge of the box, while black dots represent outliers. The right panel shows mean SLOPE values by WORD among only those participants who had the expected response pattern. Error bars represent ± one standard deviation. The y-axis is inverted, such that higher (more negative) values correspond with steeper slopes. The left three words in each panel represent [a]/[ɒ]$_A$, while the right three words in each panel represent [ɔ]/[ɒ]$_O$. 246

Figure 5.20: Number of participants with the expected response pattern by WORD. The word *pat* had the highest number of expected patterns (N=32). The left three words represent [a]/[ɒ]$_A$, while the right three words represent [ɔ]/[ɒ]$_O$. 251

Figure 5.21: Violin plots of BOUNDARY LOCATION by WORD with overlaid boxplots. These plots include those participants who were deemed to have either the expected or opposite response pattern. Values toward the left of the plot correspond with boundary locations farther toward the dialect-like end of the continuum, while values toward the right of the plot correspond with boundary locations farther toward the SEA-like end of the continuum. As shown in the plot and noted in Section 5.6.1, some participants' boundary locations fell outside of the values of the continuum steps (i.e., were less than 1 or greater than 7). The box represents the range between the first and third quartiles (interquartile range), while the horizontal line represents the median. The whiskers represent data points that lie up to 1.5 times the interquartile range from the edge of the box, while black dots represent outliers. The bottom three words represent [a]/[ɒ]$_A$, while the top three words represent [ɔ]/[ɒ]$_O$. 254

Figure 6.1: Violin plots showing the relationship between ACCURACY in the AX task (calculated for each vowel pair) and RESPONSE PATTERN in the dialect categorization task (also pooled for each vowel pair). The plots exclude Participant 36, who was excluded from all analyses involving the AX task. The box represents the range between the first and third quartiles (interquartile range), while the horizontal line represents the median. The whiskers represent data points that lie up to 1.5 times the interquartile range from the edge of the box, while black dots represent outliers. 272

Figure 6.2: Violin plots showing the relationship between *d'* in the AX task (calculated from pooled data for each vowel pair) and RESPONSE PATTERN in the dialect categorization task (also pooled for each vowel pair). The plots exclude Participant 36, who was excluded from all analyses involving the AX task. The box represents the range between the first and third quartiles (interquartile range), while the horizontal line represents the median. The whiskers represent data points that lie up to 1.5 times the interquartile range from the edge of the box, while black dots represent outliers. 273

Figure 6.3: Relationship between BIRTH YEAR and *d'* in the AX task. Each dot represents one participant's *d'* value for one individual word. The *d'* values are grouped by vowel pair. The overlaid regression lines have a negative slope for both vowel pairs, indicating that younger people had lower *d'* values (i.e., lower accuracy on the AX task). 274

Figure 6.4: Relationship between BIRTH YEAR and SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED from the AX task (in Bark). Each dot represents one participant's interval value for one individual word. The values are grouped by vowel pair. The slopes of the overlaid regression lines indicate that the general age effect (younger people having larger intervals [i.e., worse discrimination]) found in pooled models in Chapter 4 was driven by [a]/[ɒ]$_A$ more so than by [ɔ]/[ɒ]$_O$, as the former has a more strongly positive slope. 275

Figure 6.5: Proportion of correct responses by EDUCATION LEVEL for AX varying intervals for [a]/[ɒ]$_A$. Purple represents correct responses, while yellow represents incorrect responses (i.e. failure to discriminate the interval). 275

Figure 6.6: Violin plots showing the relationship between BIRTH YEAR and RESPONSE PATTERN (expected vs. non-expected) in the dialect categorization task. The box represents the range between the first and third quartiles (interquartile range), while the horizontal line represents

the median. The whiskers represent data points that lie up to 1.5 times the interquartile range from the edge of the box, while black dots represent outliers. The effect of BIRTH YEAR on RESPONSE PATTERN was sub-significant (p=0.079) in models that pooled both vowel pairs without interactions (presented in Chapter 5), such that older people had more expected response patterns. 276

Figure 6.7: Violin plots showing the relationship between BIRTH YEAR and RESPONSE PATTERN (expected vs. non-expected) in the dialect categorization task separately for each vowel pair. The box represents the range between the first and third quartiles (interquartile range), while the horizontal line represents the median. The whiskers represent data points that lie up to 1.5 times the interquartile range from the edge of the box, while black dots represent outliers. The plots indicate that the weak general effect of BIRTH YEAR shown in **Figure 6.6** above is driven primarily by [ɔ]/[ɒ]o. 277

Figure 6.8: Violin plots showing the relationship between GENDER and SLOPE in the dialect categorization task, among only those participants who showed the expected response pattern. The box represents the range between the first and third quartiles (interquartile range), while the horizontal line represents the median. The whiskers represent data points that lie up to 1.5 times the interquartile range from the edge of the box, while black dots represent outliers. The y-axis is inverted, such that higher (more negative) values correspond with steeper slopes. The plots show that women had steeper (more negative) slopes, indicating more categorical response patterns. 277

Figure 6.9: Process of categorizing participants based on their responses to the AX and dialect categorization tasks. This process used the RESPONSE PATTERN, SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED and BOUNDARY LOCATION metrics described in Chapters 4 and 5..... 280

Figure 6.10: Groupings and sub-groupings of participants resulting from the process of categorizing participants based on their responses to the AX and dialect categorization tasks. This process used the RESPONSE PATTERN, SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED and BOUNDARY LOCATION metrics described in Chapters 4 and 5..... 281

Figure 6.11: Amount of participants who did not show a categorical RESPONSE PATTERN (either in the expected or opposite direction) for each word in the dialect categorization task. Participants in the "good_AX" column reached consistent 75% accuracy for some interval in the varying intervals portion of the AX task, but nevertheless did not have a categorical response for the given word in the dialect categorization task. Participants in the "bad_AX" column did not reach 75% accuracy at any interval in the varying intervals portion of the AX task for the given word. 284

Figure 6.12: Amount of participants for each word who showed a categorical RESPONSE PATTERN (either in the expected or opposite direction) in the dialect categorization task. Participants are sorted according to the relationship between their BOUNDARY LOCATION from the categorization task and their responses to the same word in the AX task. The yellow column represents participants who had a categorical response despite never reaching 75% consistent accuracy in the varying intervals portion of the AX task. The bright and pale pink

columns represent participants whose BOUNDARY LOCATION was to the left or right of their SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED from step 7 in the varying intervals portion of the AX task. The purple column represents participants whose BOUNDARY LOCATION and SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED corresponded. 285

Figure 6.13: Example of participant (Participant 13; *pat*) whose **BOUNDARY LOCATION corresponded with the SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED from step 7 in the AX varying intervals portion**. The SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED was 3-7, and the participant's BOUNDARY LOCATION was 3.5. The left graph is the participant's responses to the dialect categorization task (purple (1) = "Gavaṛ"; yellow (0) = "not Gavaṛ"). The graph on the right represents the participant's responses to AX varying intervals (right column = correct response; left column = incorrect response; colors represent different intervals). 286

Figure 6.14: Comparison of BOUNDARY LOCATION with areas of relatively high accuracy on the consistent intervals portion of the AX task (among those participants who **never reached** consistent 75% accuracy on the AX varying intervals portion for the relevant word). 287

Figure 6.15: Example of participant (Participant 24; *pat*) who **had a perceptual boundary in the dialect categorization task despite never reaching consistent 75% accuracy in the varying intervals portion of the AX task. The BOUNDARY LOCATION is also not explainable by AX accuracy elsewhere along the continuum** (the participant had 0% accuracy for all 2-step intervals). The upper-left graph is the participant's responses to the dialect categorization task (purple (1) = "Gavaṛ"; yellow (0) = "not Gavaṛ"). The lower-left graph represents the participant's accuracy on AX consistent intervals (dark blue=correct; green=incorrect). The graph on the right represents the participant's responses to AX varying intervals (right column = correct response; left column = incorrect response; colors represent different intervals). The participant's BOUNDARY LOCATION was 2.75. 288

Figure 6.16: Example of participant (Participant 20; *mas*) who **had a perceptual boundary in the dialect categorization task despite never reaching consistent 75% accuracy in the varying intervals portion of the AX task. The BOUNDARY LOCATION *is* explainable by AX accuracy elsewhere along the continuum.** The upper-left graph is the participant's responses to the dialect categorization task (purple (1) = "Gavaṛ"; yellow (0) = "not Gavaṛ"). The lower-left graph represents the participant's accuracy on AX consistent intervals (dark blue=correct; green=incorrect). The graph on the right represents the participant's responses to AX varying intervals (right column = correct response; left column = incorrect response; colors represent different intervals). The participant's BOUNDARY LOCATION was 2.5..... 288

Figure 6.17: Comparison of BOUNDARY LOCATION with areas of relatively high accuracy on the consistent intervals portion of the AX task (among those participants who **reached** consistent 75% accuracy on the varying intervals portion for the relevant word, and whose BOUNDARY LOCATION and SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED did not correspond). 289

Figure 6.18: Example of participant (Participant 24; *mas*) who **had a perceptual boundary in the dialect categorization task and reached consistent 75% accuracy in the varying**

intervals portion of the AX task, but whose BOUNDARY LOCATION falls to the right of the SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED from step 7. The SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED was 1-7, while the BOUNDARY LOCATION was 4.76. The boundary location is also not explainable by AX accuracy elsewhere along the continuum, as the participant had 0% accuracy at the interval 5-7**.** The upper-left graph is the participant's responses to the dialect categorization task (purple (1) = "Gavaṛ"; yellow (0) = "not Gavaṛ"). The lower-left graph represents the participant's accuracy on AX consistent intervals (dark blue=correct; green=incorrect). The graph on the right represents the participant's responses to AX varying intervals (right column = correct response; left column = incorrect response; colors represent different intervals). 290

Figure 6.19: Comparison of individual participants' produced differences in F2 (in raw Bark values) with Kewley-Port's (2001) discriminability threshold of 0.5 Bark. The threshold is represented by the horizontal dotted line, while the two columns represent the difference in F2 between the members of each vowel pair in each participant's production data. Participants are ordered by their value of F2 for [a]/[ɒ]$_A$, from highest to lowest. 297

Figure 6.20: Individual Pillai scores by GENDER for **[a]/[ɒ]$_A$** with superimposed distinction thresholds. Participants are ordered from lowest Pillai score (most overlapped vowels) to highest Pillai score (least overlapped vowels). 298

Figure 6.21: Individual Pillai scores by GENDER for **[ɔ]/[ɒ]$_O$** with superimposed distinction thresholds. Participants are ordered from lowest Pillai score (most overlapped vowels) to highest Pillai score (least overlapped vowels). 299

Figure 6.22: Individual perception results for [a]/[ɒ]$_A$ of Participant 39, who had the largest Pillai score for this vowel pair. The top row represents AX varying intervals (left column = incorrect response; right column = correct response, i.e., successful discrimination). The middle row represents AX consistent intervals (dark blue (1) = correct response, i.e., successful discrimination; green (0) = incorrect response). The bottom row represents the dialect categorization task (purple (1) = "Gavaṛ"; yellow (0) = "not Gavaṛ"). 308

Figure 6.23: Individual perception results for [a]/[ɒ]$_A$ of Participant 35, who had the second largest Pillai score for this vowel pair. The top row represents AX varying intervals (left column = incorrect response; right column = correct response, i.e., successful discrimination). The middle row represents AX consistent intervals (dark blue (1) = correct response, i.e., successful discrimination; green (0) = incorrect response). The bottom row represents the dialect categorization task (purple (1) = "Gavaṛ"; yellow (0) = "not Gavaṛ"). 309

Appendix Figure E.1: Individual responses for *pat* 357

Appendix Figure E.2: Individual responses for *mas* 358

Appendix Figure E.3: Individual responses for *mard* 358

Appendix Figure E.4: Individual responses for *pos* 359

Appendix Figure E.5: Individual responses for *tsots*..... 359

Appendix Figure E.6: Individual responses for *xoz*..... 360

Appendix Figure F.1: Dialect categorization boundary locations by word for the [a]/[ɒ]$_A$ vowel pair. The color coding indicates participants who had response patterns with expected directionality (i.e., categorizing the dialect-like end of the continuum as dialectal and the SEA-like end of the continuum as non-dialectal) vs. those whose response pattern had the opposite directionality. Each dot represents one participant. 361

Appendix Figure F.2: Dialect categorization boundary locations for the [a]/[ɒ]$_A$ vowel pair (x-axis) in relation to ED-median (from non-Bark-converted data) from participants' productions of the same vowel pair (y-axis) with overlaid regression lines. The color coding indicates participants who had response patterns with expected directionality (i.e., categorizing the dialect-like end of the continuum as dialectal and the SEA-like end of the continuum as non-dialectal) vs. those whose response pattern had the opposite directionality. Each dot represents one participant's boundary location for one word. 361

Appendix Figure F.3: Dialect categorization boundary locations for the [a]/[ɒ]$_A$ vowel pair (x-axis) in relation to Pillai scores from participants' productions of the same vowel pair (y-axis) with overlaid regression lines. The color coding indicates participants who had response patterns with expected directionality (i.e., categorizing the dialect-like end of the continuum as dialectal and the SEA-like end of the continuum as non-dialectal) vs. those whose response pattern had the opposite directionality. Each dot represents one participant's boundary location for one word. 362

List of Appendices

Appendix A: Word List 336

Appendix B: Demographic and Language Background Questionnaire 341

Appendix C: Consent Forms 344

Appendix D: Verbal Task Instructions 354

Appendix E: Individual Responses to Dialect Categorization Task 357

Appendix F: Additional Figures Related to BOUNDARY LOCATION 361

Abstract

This dissertation explores production and perception of dialect vowels in Gavaṛ, Armenia, focusing on how Gavaṛ residents produce and perceive differences between the members of two specific pairs of vowels. One of these pairs ([a]/[ɒ]$_A$) is allophonic and socially salient, while the other ([ɔ]/[ɒ]$_O$) is contrastive and less salient. In exploring the production and perception of these vowels, the dissertation seeks to answer two broad questions emerging from previous phonetic and sociophonetic research. The first question asks how social and phonological factors influence production, perception, and production-perception links, while the second explores implications of the study's results for what it means for vowels to be distinct from one another, including implications for definitions of concepts such as "merger." The dissertation presents production data and the results of a two-part perception experiment, which consisted of an AX discrimination task and a dialect categorization task (in which participants heard words with varying vowels and categorized them as characteristic or not characteristic of Gavaṛ's dialect). Both tasks presented stimuli including vowel continua ranging from dialect-like pronunciations ([ɒ]$_A$ and [ɒ]$_O$) to Standard Eastern Armenian pronunciations ([a] and [ɔ]).

Evidence was found for the influence of both social and phonological factors on production and perception, though findings related to production-perception links were more limited, as clear links were generally not found. The most notable finding related to the first research question was the evidence from the perception experiment for the influence of both phonological and social factors on perceptual outcomes. In particular, the socially salient, allophonic [a]/[ɒ]$_A$ pair showed relatively poor discrimination in the AX task, especially given the large acoustic differences presented in its stimuli. This result aligns with previous literature according to which allophonic differences are generally difficult to discriminate. Nevertheless, this pair was categorized more accurately in the dialect categorization task. This contrast between the results of the two tasks suggests that social factors (such as salience of vowels and social framing of the task) can facilitate the perception of difficult-to-discriminate allophonic differences. Additionally, the results of the perception tasks suggested that different areas of the

[a]/[ɒ]$_{A}$ continuum are perceived in different ways. In particular, small differences from the standard variant [a] are potentially perceived as coarticulatory differences (and thus discriminated with the greatest difficulty), while larger differences are more likely perceived as Gavaṛ's socially salient dialectal allophone. In relation to the second research question, the results suggested that concepts such as "merger" and "near-merger" of vowels should not be defined in a context-independent way, as acoustic differences of the same magnitude can be perceived differently in different language varieties and speech contexts according to the phonological and social characteristics of the vowels in question.

In addition to its theoretical contributions, this dissertation contributes to the documentation of the current state of regional dialects of Armenian, which are understudied especially as it pertains to interactions with supralocal varieties and sociolinguistic variation in the maintenance of dialect features. In documenting contemporary urban speech, variation in the production of dialect features, and Gavaṛ residents' perception of such variation, this dissertation draws attention to types of speech that have been relatively overlooked in previous research on Armenian, highlighting the complexities emerging from interactions between dialects and supralocal varieties.

Chapter 1 Introduction

1.1 Chapter summary

This chapter presents the broad research questions explored in this dissertation, while also summarizing the components of the dissertation and how they seek to address these questions. The chapter then provides background information about the context in which the research was carried out, discussing the sociolinguistic environment of the Republic of Armenia, the Armenian dialectological tradition, and the current state of regional dialects in Armenia. Background information about the Bayazet dialect (spoken in and around the city of Gavaṛ,[1] Armenia) is then presented, including an overview of previous research on the dialect and of its vowel system, the production and perception of which is explored in this dissertation. Finally, the chapter provides an overview of how the broad research questions are motivated by previous literature, as well as an outline of the remaining chapters of the dissertation.

1.2 Research questions and components of dissertation

In examining the production and perception of vowels in Gavaṛ, Armenia, this dissertation explores two overarching questions:

1) How do social and phonological factors influence production and perception of vowels, as well as the relationship between production and perception?

[1] Except for names that appear frequently in English (e.g., "Yerevan") and names that appear in Latin letters in the original source, Armenian names, along with the titles of works and transliterated words in quotes from works in Armenian, are transliterated using a modified version of the Library of Congress romanization system. Voiceless aspirated consonants have been indicated with a superscript *h* rather than an apostrophe to avoid confusion with ejective consonants. In contrast, words elicited from or produced by participants in the study are transliterated using IPA to convey dialectal sounds that cannot be transcribed using available romanization systems. Though IPA transcriptions of the six words that made up the perception stimuli are provided in Chapters 4-5, simplified transliterations (*pat, mas, mard, pos, tsots, xoz*) have generally been used in place of the full IPA in order to reduce the need for special characters (especially in figures). Simplified transliteration has also been used for the place name "Hats^{h}aṛat" in cases when it refers to a predictor in statistical models (and is thus formatted in small caps).

2) Considering that both perception and production of vowels can be influenced by social and phonological factors, what does it mean for two vowel sounds to be distinct from one another in a given language variety?

These questions have been explored through the following two phases of data collection:

1) A production study consisting of sociolinguistic interviews (including conversational speech, a picture-naming task, and a word list) carried out in July of 2022

2) A two-part perception experiment with accompanying production data (word list) carried out in June of 2025

The results of the sociolinguistic interviews carried out in 2022 not only provided insights into social variation in the maintenance of dialect vowels in Gavaṛ, but also motivated the 2025 perception experiment and the exploration of the questions listed above (see Portugal [2025] for full results of the 2022 study and Chapter 3 for detailed discussion of the motivations of the perception experiment). In particular, patterns of variation found in participants' vowel systems raised questions about how the phonology of Gavaṛ's dialect might be contributing to the observed patterns and about how it also might contribute to participants' perception of the given vowels.

The 2025 study consisted of three components (elicitation of production data and two perception tasks), which participants completed in the following order:

1) Recorded word list containing the vowels of interest

2) An AX discrimination task, in which participants were asked to discriminate differences between vowels that varied along continua from dialect pronunciations to Standard Eastern Armenian pronunciations

3) A dialect categorization task, in which participants categorized stimuli with vowels varying along continua (many of which were identical to those used in the AX task) in terms of whether they thought they were characteristic of Gavaṛ's dialect. Importantly, this task involved categorizing different pronunciations of a given lexical item (no minimal pairs were involved,

even in the case of vowel differences that are contrastive), such that the task explored social rather than phonemic categorization of vowel continua.

The results of the 2022 sociolinguistic interviews, together with the 2025 production data, contribute to this dissertation's broader goals mainly by providing information about the influence of social factors on speech production. The 2025 perception experiment, in examining vowel differences that differ in their phonological characteristics (i.e., allophonic vs. contrastive differences) and social salience, contributes to the exploration of how phonological and social factors influence perception. Beyond the differing characteristics of the vowels examined, the implementation of multiple perception tasks (with vs. without social prompting) also provides insights into the influence of social factors on perception. In comparing the results of the perception tasks with participants' productions of the relevant vowels, the dissertation investigates relationships between perception and production, including how they might be affected by task type and by the characteristics of the vowels involved. As the vowel differences under investigation are produced with a great deal of variation and may be progressing toward a more merged state (based on the 2022 production data), the investigation of their perception and production contributes to the exploration of what it means for vowels to be distinct, with consideration for individual variation and for phonological and social factors that affect both perception and production.

In addition to addressing the research questions listed above, this dissertation also seeks to contribute to the documentation of the current state of regional Armenian dialects (see further discussion in Section 1.3.1.2). Though some authors (e.g., Hodgson 2019) have observed that the use of regional Armenian dialects may be declining, there has been little empirical documentation of the current state of these dialects and of variation among speakers, despite the existence of recently published dialectological works (e.g., Katvalyan 2025). In general, despite the long tradition of Armenian dialectology, Armenian dialects have seldom been studied using modern empirical/quantitative linguistic methods. This dissertation thus builds upon the legacy of Armenian dialectology in a novel way by using dialectological works as the basis for modern empirical and experimental studies.

1.3 Background information – sociolinguistic context and vowel system of Gavaṛ

Section 1.3.1 below provides background information about the sociolinguistic context of the Republic of Armenia and Armenian dialectology, including what is known about the current state of regional dialects and about interactions between different language varieties in Armenia today. Section 1.3.2 provides information about previous research on the dialect of Armenian spoken in Gavaṛ (Bayazet dialect) and the vowel system explored in this dissertation.

1.3.1 Sociolinguistic context – Armenian language and dialectology

1.3.1.1 Sociolinguistic environment of the Republic of Armenia – tension between diglossia and prescriptivism

Armenian is a pluricentric language representing an independent branch of the Indo-European family. The two modern standards – known as Eastern and Western Armenian – emerged in the 19th century in the Russian and Ottoman empires respectively (Cowe 1992; Dum-Tragut 2009; Karapetian 2014; Oshagan 1997). Standard Eastern Armenian (SEA) is the official language of the Republic of Armenia (the successor of Soviet Armenia) (Dum-Tragut 2009: 6), where it coexists with colloquial varieties and regional dialects in a diglossic situation (Karapetian 2014: 31-33). The term "diglossia" refers to situations in which "high" and "low" language varieties are used by the same speech community in clearly separated contexts (Ferguson 1959). SEA has traditionally been used in formal contexts such as news broadcasts and school classrooms, and the linguistic gap between SEA and modern colloquial varieties has been described as considerable (Dum-Tragut 2009). However, Karapetian (2014: 32-33) has observed that the relationship between SEA and colloquial varieties is currently being renegotiated, such that colloquial language is increasingly being used in media. This tendency may be part of a more general linguistic upheaval connected with the collapse of the Soviet Union; Auer (2005: 31-32) noted that "destandardisation" processes were ongoing in post-Soviet states such as Russia, Belarus, and Ukraine.

Linguistic purism/prescriptivism has been a prominent element of language policy in post-Soviet Armenia (Abrahamian 1998; Abrahamian 2006; Dum-Tragut 2009; Dum-Tragut 2013; Portugal & Nonnenmacher 2024; Zakarian 1996). However, because of the diglossic situation, it is unclear if such ideologies have actually exerted any significant influence in terms of bringing colloquial varieties closer to SEA, with some studies (e.g., Ghazaryan 2020)

suggesting that frequently targeted elements such as foreign words are still abundant in colloquial speech.

This tension between purist/prescriptivist ideologies and the diglossic reality has also been noted by prominent Armenian dialectologist Viktor Katvalyan (2009: 3), who describes the existence of three manifestations of the Armenian language – literary (*grakan*) language, folk-colloquial (*zhoghovrdakhosakts^{h}akan*) language, and dialects – and their differing spheres of usage. Katvalyan (2009: 3) notes that the literary language is used in spheres such as state governance, politics, science, and education, while he defines folk-colloquial language as language that "approaches the literary, but also has considerable deviations from the literary norm and contains dialectal elements that are comprehensible to everyone, without being the same as any dialect,"[2] and notes that it is used in "non-official, everyday situations, in non-dialect-speaking environments."[3] Though Katvalyan (2009: 5) defends the importance of all three forms of language, he acknowledges that this approach is not universal in Armenian society, noting that "care for the literary language is sometimes manifested as swinging a sword at dialects and colloquial language,"[4] and arguing that "one should not despise or consider unacceptable the existence of colloquial language and dialects for the sake of the purity and regularity of the literary language."[5]

Though the metalinguistic commentary expressed by participants of the 2022 sociolinguistic interviews conducted in Gavaṛ is not a main topic of this dissertation (see Portugal [2025] for further discussion of this commentary), it provides much-needed empirical support for previous authors' observations about Armenia's linguistic-ideological environment. This commentary was generally consistent with what one might expect in a diglossic environment, while also sometimes alluding to the tension and prescriptivism noted by Katvalyan (2009). The commentary demonstrated that participants did not unambiguously understand SEA as being the most prestigious or "correct" language variety, with types of overt

[2] "... գրականին մոտեցող, բայց և գրական նորմայից զգալի շեղումներ ունեցող, բոլորին հասկանալի բարբառային տարրեր պարունակող, բայց որևէ բարբառի հետ չնույնացող…"
[3] "... ոչ պաշտոնական, առօրյա իրադրություններում, ոչ բարբառախոս միջավայրում:"
[4] "... գրական լեզվի նկատմամբ հոգատարությունը երբեմն դրսևորվում է բարբառների ու խոսակցական լեզվի դեմ սուր ճոճելով…"
[5] "... հանուն գրական լեզվի մաքրության, կանոնականության՝ չպետք է ընդհանրապես արհամարհել ու մերժելի համարել խոսակցական լեզվի ու բարբառների գոյությունը…"

prestige being assigned both to SEA and to dialects. Despite generally expressing positive attitudes toward their dialect or dialects in general, participants also noted attempting to speak closer to SEA in certain contexts, often stressing the importance of having a means of interdialectal communication. Participants also associated using SEA with contexts such as educational and state institutions, with one participant explicitly stating that one should speak dialect in everyday life. Thus, the commentary suggested that participants believe different language varieties should be used in different contexts. Together with the general findings of the 2022 study (according to which many speakers were maintaining at least some dialect vowels), this in turn suggests that linguistic prescriptivism and purism have had limited practical effects on everyday speech and speakers, being counteracted to some extent by Armenia's diglossic environment, which seems to prevent speakers from feeling that they should speak SEA in all situations. Nevertheless, commentary that implied SEA was more "correct" or "cultured" was present, and some participants also mentioned avoiding dialect features when speaking with their own young children, despite expressing positive attitudes toward the dialect in other parts of their interviews. Thus, while participants' metalinguistic commentary demonstrated some qualities that may be specific to diglossic situations, the common tendency of paying "lip-service to correctness and prestige variants" (Milroy and Milroy 2012: 16) was also noted.

1.3.1.2 Armenian dialectology, the current state of dialects, and interactions with supralocal varieties

Armenian dialectology began to develop in the 19th century (Weitenberg 2002: 143), with Hrach^{h}ya Acharyan being considered the founder of modern Armenian dialectology (Weitenberg 2002: 144-145; see e.g., Acharyan 1911). Armenia's dialectological tradition continued to develop throughout the 20th century, culminating in Jahukyan's (1972) *Statistical dialectology*, considered the most comprehensive source on Armenian dialects (see Weitenberg 2002: 148). In the late 20th century, an Armenian dialect atlas project was also initiated, with guidelines for data collection being published by the Language Institute of the Armenian National Academy of Sciences in 1977 (Hrach^{h}ya Acharyani anvan Lezvi institut 1977), and several volumes being published based on this data (e.g., Galstyan 2010; Katvalyan & Abrahamyan 2016; see also preface of Katvalyan [2025] for information about this project). Armenian dialectologists have continued carrying out fieldwork in the 21st century, with

volumes such as Katvalyan (2018) and Katvalyan (2025) presenting the results of relatively recent fieldwork.

In spite of this long and continuing tradition, the contemporary state of regional dialects in Armenia remains understudied and poorly understood, due in part to the limited methods of dialectological research and the lack of transparent presentation of their methodologies and results (see Section 1.3.2.1 below for further discussion of these issues). Though there were once as many as 120 dialects of Armenian (Jahukyan 1972: 32), some of which are mutually unintelligible (Vaux 1998: 7), most western dialects were lost after the Armenian Genocide (Manoukian 2023; Weitenberg 2002: 142). Several authors have noted the potential instability of the remaining eastern dialects in the Republic of Armenia, particularly in urban areas. For example, Weitenberg (2002: 142) notes that "dialects are still spoken in rural areas; the influence of the literary language is penetrating everywhere through schools and mass communication." Hodgson (2019: 95-96) writes that a "quasi-standard colloquial language" based on Yerevan's colloquial variety is used in most urban areas, while regional dialects remain in rural areas.

One question that emerges from Weitenberg and Hodgson's observations (together with Karapetian's [2014] observations about potential ongoing renegotiation of the diglossic norm) is whether the literary language (i.e., SEA) or the colloquial variety of Yerevan plays a more prominent role in any potential ongoing leveling of regional dialects. The analysis of vowels presented in Portugal (2025) could not answer this question, as SEA and the colloquial variety of Yerevan use the same vowel system (see Markosyan 1989; Seyfarth et al. 2023). Nevertheless, participant commentary from the 2022 sociolinguistic interviews sheds some light on this question. In this commentary (which comes from the interviews analyzed in Portugal [2025], but which is not extensively discussed in that publication), many participants expressed disdain for the colloquial language of Yerevan or denied its influence on their own speech, while stating that they attempt to speak closer to the "literary" language (i.e., SEA) in some situations. However, some degree of confusion about terminology was apparent in this commentary. For example, many of the speakers who opposed the use of the colloquial variety of Yerevan in media or in their own speech referenced "jargon" (*zhargon,* i.e., 'slang') or "street jargon" in their opposition to the use of this variety, making it difficult to understand if they dislike the variety in general or only certain subsets of its vocabulary. Similarly, it was sometimes unclear whether participants understood the term "standard (literary) language" to refer to SEA or to a "quasi-standard

colloquial language" of the sort mentioned by Hodgson (2019: 96) (which is conceptually similar to Katvalyan's [2009] "folk-colloquial" language). One participant mentioned that many people mistakenly consider elements of Yerevan's dialect to be SEA. Relatedly, many participants who claimed to attempt to speak closer to SEA with people from other parts of Armenia and/or expressed a preference for the use of SEA in media used widely recognized elements of Yerevan's colloquial language, such as *vɔnt͡sʰ* ('how') and *χi* ('why') during their interviews and more generally while conversing with the research team.

Taken collectively, these factors suggest that a "quasi-standard colloquial language" may in fact be more relevant in ongoing processes of dialect leveling than SEA, whether speakers themselves are conscious of this or not. Indeed, Hodgson (2019: 95) also makes an observation along these lines, writing "Some of the influence of the 'standard' language on local dialects is in fact from this form, rather than from literary SEA." This also follows logically from Armenia's history of diglossia, as in this scenario, the spoken dialects would be yielding to another spoken variety, rather than to a variety that is not used by any group of speakers in daily life. Despite some apparent prestige being attached to the colloquial language of Yerevan, as suggested by its generally widening scope of usage, the aforementioned metalinguistic commentary expressing disdain for this variety suggests that the idea of a "standard" or "literary" language that is separate from the language of the capital city still holds appeal for many speakers. This point of tension highlights the ongoing competition that exists between different supralocal varieties in relation to regional dialects in Armenia today.

Despite previous authors' observations about potential dialect leveling, along with some evidence from the 2022 sociolinguistic interviews in support of their observations, systematic investigation of dialect maintenance and of sociolinguistic variation in the use of dialects in Armenia is nevertheless lacking. The post-Soviet region is generally underrepresented in quantitative and variationist sociolinguistic research (see Portugal 2025), and in Armenia in particular, there is a lack of quantitative studies of recent changes that regional Armenian dialects have potentially undergone under the pressure of SEA and the colloquial language of Yerevan. Portugal (2025) began to address this gap by studying sociolinguistic variation in the maintenance of dialect vowels in Gavaṛ, and this dissertation contributes additional information about the current state of Gavaṛ's dialect by providing perception data and additional production data.

1.3.2 Bayazet dialect

1.3.2.1 Overview of previous research

Both phases of data collection mentioned above were carried out in Gavaṛ, Armenia. Gavaṛ is the administrative center of Geghark^{h}unik^{h} province and has a permanent population of 17,741 according to the 2022 census (Statistical committee of the Republic of Armenia 2022). The vowel system of the dialect spoken in Gavaṛ is ideal for a case study of interactions between social and phonological factors, as it contains several additional phonemic and allophonic vowel distinctions not found in SEA, some of which are socially salient (see Section 1.3.2.2 below). The local dialect is often referred to as the "Bayazet dialect" in the literature, as the residents of Gavaṛ and the surrounding villages are said to descend primarily from residents of the city of Bayazet (located in present-day Turkey) and its surrounding villages, who migrated en masse around 1830 when Bayazet was ceded to the Ottomans by the Russian Empire (Katvalyan 2016: 31). All participants of both phases of data collection were born in Gavaṛ, and most had also spent their youth through age 18 there. Both phases of data collection were also carried out in collaboration with local research assistants, who conducted interviews, explained experiment protocols, and aided in the development of elicitation materials.

Gavaṛ was selected as a suitable site for exploring variation in the maintenance of regional dialects not only because it is an urban settlement (where dialect leveling may be occurring according to the observations of Weitenberg [2002] and Hodgson [2019]), but also because its dialect has been relatively well documented by Armenian dialectologists, allowing for the comparison of new data with previous dialectological descriptions. Despite the limitations of the information presented in such dialectological works, this dissertation takes the same approach adopted by Maguire (2007) in considering traditional dialectology to be a valuable source of information that can be examined alongside other types of data, provided that one is cautious in interpreting the scope of traditional dialect data (see Maguire 2007: 90).

Acharyan (1911: 37-46) provides a brief description of the Bayazet dialect, which he classifies as a sub-branch of the Yerevan dialect. A more extensive description of the dialect's features can be found in Baghramyan (1972: 11-34), though it is notable that both Acharyan's and Baghramyan's descriptions differ from later works (e.g., Katvalyan 2016) in significant ways, either not mentioning certain features or describing the pronunciations and phonological environments of certain sounds differently than later works (see further discussion in Section

1.3.2.2 below). According to Katvalyan (2016: 28), Jahukyan was the first to identify the Bayazet dialect as a separate dialect, rather than a sub-dialect of Yerevan's (see Katvalyan [2016] for a comprehensive overview of the history of documentation of the Bayazet dialect). Many dialectological works related to the Bayazet dialect have been published in the 21st century, including a collection of stories in the dialect (Katvalyan 2009), articles about specific features (Katvalyan 2012; Mkrtch^{h}yan 2015, 2016), a comprehensive dialect grammar (Katvalyan 2016), a dialect dictionary (Katvalyan 2021), volumes describing all the dialects of Geghark^{h}unik^{h} province (Katvalyan 2018; Katvalyan et al. 2021), and a dialect atlas of Geghark^{h}unik^{h} province (Katvalyan 2025).

Despite this extensive list of works, it is difficult to understand the extent to which these dialectological descriptions reflect the contemporary speech of Gavaṛ's population in general. Authors such as Katvalyan (2016, 2018) and Mkrtch^{h}yan (2015) maintain that dialect speakers in Gavaṛ have preserved regional features. In particular, Katvalyan (2016: 19) states that there are not phonetic and grammatical differences between the Bayazet dialect of the late 19th and early 20th centuries and the dialect of today (2016: 19), while Katvalyan (2018: 12) elaborates, explaining that "in the settlements of Geghark^{h}unik^{h} province, the main phonetic and grammatical features of dialect speech remain stable, and there can be no discussion of dialects disappearing under the influence of the literary language in this region. Dialects develop according to their regular course, naturally having shifts especially in the vocabulary."[6] While these statements seem to imply that maintenance of dialects entails relative stability at least of phonetic and grammatical features, a more extensive discussion from Katvalyan (2009: 5-6) seems to argue for a less restrictive definition of "dialect" and of dialect maintenance:

> For decades, many linguists have mournfully announced the death of dialects, as the result of widespread education and enlightenment and a consequence of the literary language's victory march, confusing dialects' course of development with disappearance. In reality, within the framework of language in general, dialects also undergo changes, having especially noticeable shifts in vocabulary… . Over time, dialects can also undergo

[6] "... Գեղարքունիքի մարզի բնակավայրերում բարբառային խոսքը հաստատուն է պահում հնչյունական եւ քերականական հիմնական հատկանիշները, եւ այս տարածաշրջանում գրական լեզվի ազդեցությամբ բարբառների վերանալու մասին խոսք լինել չի կարող: Բարբառները զարգանում են բնականոն հունով՝ բնականաբար տեղաշարժեր ունենալով հատկապես բառապաշարում:"

> changes in linguistic features. But even so, all of this cannot mean the gradual disappearance of dialects, because the phenomenon of "dialect" is characterized not so much by this or that layer of dialect words or by this or that dialect feature (which undergo various changes over time), but rather by the unrestrictedness of speech, the non-uniformity of forms, non-cultivatedness, features characteristic only of spoken language, the combinations of dialect features characteristic only of the given area… . Widespread command of the literary language and the ability of many or all to communicate in it when necessary cannot become a reason for the loss of dialects.[7]

Thus, while it seems that Katvalyan acknowledges that language varieties change over time and understands the concept of "dialect" in a way such that a dialect undergoing changes is not necessarily disappearing, the lack of transparency about methodology in dialectological sources nevertheless creates difficulty for the interpretation of the presented dialect descriptions. In general, the sources of data in Armenian dialectological works are usually not transparent, and the aforementioned statements about dialect maintenance in Gavaṛ/Geghark^{h}unik^{h} are not accompanied by discussion of which residents of a given area are considered dialect speakers, or of what types of residents were surveyed in dialectological fieldwork. Some sources do provide some information about data collection. For example, Katvalyan (2018: 14-15) notes that the book presents the results of audiovisual materials and surveys gathered during fieldwork, and he also includes an acknowledgements section in which the names of "dialect speakers and supporters" are listed. It is unclear if this section lists the names of all speakers who participated

[7] "Տասնամյակներ շարունակ շատ լեզվաբաններ գուժել են բարբառների մահը՝ իբրև համատարած կրթության, լուսավորության տարածման, գրական լեզվի հաղթարշավի հետևանք՝ շփոթելով բարբառների զարգացման ընթացքը վերացման հետ: Իրականում ընդհանուր լեզվի շրջանակներում բարբառներն էլ փոփոխություններ են կրում, հատկապես նկատելի են տեղաշարժերը բառապաշարում… : Ժամանակի ընթացքում բարբառներում կարող են տեղի ունենալ նաև լեզվական հատկանիշների փոփոխություններ: Բայց և այնպես, այդ ամենը չի կարող նշանակել բարբառների աստիճանական վերացում, որովհետև բարբառ իրողությունը բնութագրվում է ոչ այնքան բարբառային բառերի այս կամ այն շերտով կամ բարբառային այս կամ այն հատկանիշով (որոնք ժամանակի ընթացքում կրում են տարաբնույթ փոփոխություններ), որքան խոսքի անկաշկանդությամբ, ձևերի անմիօրինակությանբ, անմշակվածությամբ, միայն բանավոր խոսքին բնորոշ առանձնահատկություններով, բարբառային հատկանիշների՝ միայն տվյալ տարածքին բնորոշ զուգորդումներով… : Բարբառի կորստի պատճառ չի կարող դառնալ գրական լեզվին համատարած տիրապետումը, ըստ անհրաժեշտության՝ նրանով հաղորդակցվելու շատերի կամ բոլորի կարողությունը:"

in the fieldwork, but if this is in fact the case, then the number of participants is very small – for example, the section lists only 9 people from Gavaṛ (3 from Gavaṛ in general, 3 from the Hatsʰaṛat neighborhood, and 3 from the Artsvakʰar neighborhood). In a later section of the same book (Katvalyan 2018: 106), it is noted that information about dialectal features used in Gavaṛ has also been drawn from materials gathered in the abovementioned dialect atlas project in 1999. The elicitation methods of the dialect atlas are known thanks to the 1977 publication (Hrachʰya Acharyani anvan Lezvi institut 1977), and Katvalyan (2018: 106) lists information about a specific notebook kept in Armenia's Language Institute, in which the data gathered in Gavaṛ can be found. However, I visited the Language Institute in 2025 in the hopes of viewing this notebook, and it turned out that it had been lost. Accordingly, it is not possible to know how many people or what kinds of people participated in the dialect atlas's surveys in Gavaṛ. Other sources (Katvalyan et al. 2021: 3; Katvalyan 2025: 17) mention the same fieldwork mentioned by Katvalyan (2018), noting that it was carried out in 2013-2015 and consisted of "audiovisual recordings and structured interviews" (Katvalyan 2025: 17). However, these sources also do not provide information about the number of participants or about criteria used in selecting participants.

Because of this lack of transparency in providing information about participants and methods, one wonders if surveys such as the 2013-2015 fieldwork and the Armenian dialect atlas fieldwork relied on a relatively narrow definition of "dialect speaker" of the type usually implemented in traditional dialectology. In other words, it is possible that the data presented in these sources was collected from relatively conservative dialect speakers, and that such speakers were intentionally sought out to participate in surveys, as is customary in traditional dialectology (see Chambers & Trudgill 1998; Maguire 2007; see also discussion in Section 2.3.3.1). Katvalyan's statements about the phonetic and grammatical characteristics of the Bayazet dialect remaining relatively unchanged since the early 20th century contradict Baghramyan's (1972) observation of ongoing leveling of regional dialect features in the area around Lake Sevan (located in Gegharkʰunikʰ province) and in Gavaṛ specifically, especially in the speech of young people. Baghramyan writes that it is already impossible to find "pure dialect speakers" ("մաքուր բարբառախոսներ") even in remote villages (1972: 3). In describing the Bayazet dialect in particular, Baghramyan (1972: 13) writes that "both the phonetic structure and the grammar and vocabulary are changing over time, becoming more similar to the literary

language,"[8] and notes that the book primarily presents the speech of the older generation in describing dialect features. The results of the 2022 sociolinguistic interviews conducted in Gavaṛ (presented in Portugal [2025]) also contradict Katvalyan's statements, as many speakers were found to be maintaining some but not all of the dialect vowels described by Katvalyan, while others used a vowel system very close to SEA / the Yerevan dialect, and there was no participant who clearly and consistently maintained the entire dialect vowel system. Given these contradictions, it seems possible that Katvalyan's works (and Armenian dialectological works in general) generally align with the broader goals of traditional dialectology, seeking out representatives of conservative dialect speech rather than aspiring to document contemporary speech and linguistic variation across the speech community as a whole. Thus, while dialectological works are a valuable source of information and served as the foundation for the studies presented in this dissertation, their findings must be interpreted with consideration for their goals and methodologies.

1.3.2.2 Vowels

Vowels were selected as the focus of Portugal (2025) and of this dissertation because the Bayazet dialect's vowel system (as described in sources like Katvalyan [2016]) differs in substantial ways from those of SEA and of Yerevan's dialect. Accordingly, if there is variation and ongoing change in the dialect resulting from contact with these supralocal varieties, vowels are one set of features where we can expect to find such variation and change. Some aspects of the vowel system are also socially salient according to Katvalyan (2016) (and as confirmed by the 2022 interview data), which made vowels a potentially fruitful line of inquiry to be explored in the perception experiment mentioned above.

The vowel systems of the Bayazet dialect and SEA are shown below in **Figure 1.1**. IPA transcriptions of vowels of the Bayazet dialect are based on my own interpretations of Katvalyan's (2016) descriptions. In the Bayazet dialect, the rounded low back vowel [ɒ] is in complementary distribution with [a], appearing word-initially and in the first syllables of words after nasals and voiceless unaspirated consonants with the exception of [χ] (Katvalyan 2016: 81-82). Katvalyan (2016: 80) describes the [ɒ] vowel (in cases where it is an allophone of [a]) as the

[8] "... և հնչյունական կազմը և քերականությունն ու բառապաշարը գնալով վերափոխվում, ավելի են մոտենում գրական լեզվին:"

dialect vowel system's most salient aspect, and this assessment was supported by the 2022 sociolinguistic interview data. Interestingly, despite the apparent salience of this vowel, it is absent from Acharyan's (1911) brief description. In contrast, this sound is described by Baghramyan (1972: 15), though Baghramyan admits to having failed to determine all the phonological rules governing its appearance.

At the same time, according to Katvalyan (2016), a vowel with the same pronunciation as the allophone of [a] is in a contrastive relationship with [ɔ] that contains some remnants of allophony. According to Katvalyan, [ɒ] and [ɔ] represent separate phonemes, although only [ɔ] can appear before [v], as well as before [ʁ] in verbal participles (2016: 90-91). Additionally, in older layers of vocabulary, [ɒ] appears in stressed positions and [ɔ] appears in non-stressed positions (Katvalyan 2016: 89), although this relationship is not always consistent when stress changes due to the addition of inflectional and derivational suffixes (see Katvalyan 2016: 89, 178). Similarly, [ie] and [ɛ] are in a relationship that is sometimes conditioned by stress or phonetic environment (Katvalyan 2016: 92-95, 178), though minimal pairs also exist (Katvalyan 2016: 95). The vowel [æ] represents an independent phoneme and appears mostly in borrowings from other languages and dialects (Katvalyan 2016: 83-85).

Although Katvalyan (2016: 67) characterizes [ɒ] (both as an allophone of [a] and as a contrastive variant of [ɔ]) as diphthongal, transitioning from rounded to unrounded (i.e., pronounced something like [oa]), he also writes (2016: 68) that its rounding is relatively weak in Gavaṛ compared to some villages, possibly in connection with its being the regional center. As pronunciations that were perceptibly (to me) different from [a] but not audibly diphthongized were widespread among participants of the 2022 sociolinguistic interviews, the transcription [ɒ] has been used for this vowel throughout this dissertation. In later chapters of this dissertation, the transcriptions [ɒ]$_{A}$ and [ɒ]$_{O}$ are used to differentiate [ɒ] as an allophone of [a] and [ɒ] as a variant of [ɔ]. Though Katvalyan (2016) characterizes [ɒ]$_{A}$ and [ɒ]$_{O}$ as being pronounced identically, some other authors' descriptions are not consistent with this. For example, Baghramyan (1972: 14) writes that [ɒ]$_{O}$ is pronounced [ue] or [uo], while Khachaturian (1992: 118) writes that the two vowels are "almost identical… though some subtle difference can be detected between them, which however is not always present." The transcription [ie] has been used in this dissertation despite diphthongized pronunciations also being rare for this word class in the 2022 interview data, as the few instances that were audibly different from [ɛ] were strongly diphthongized.

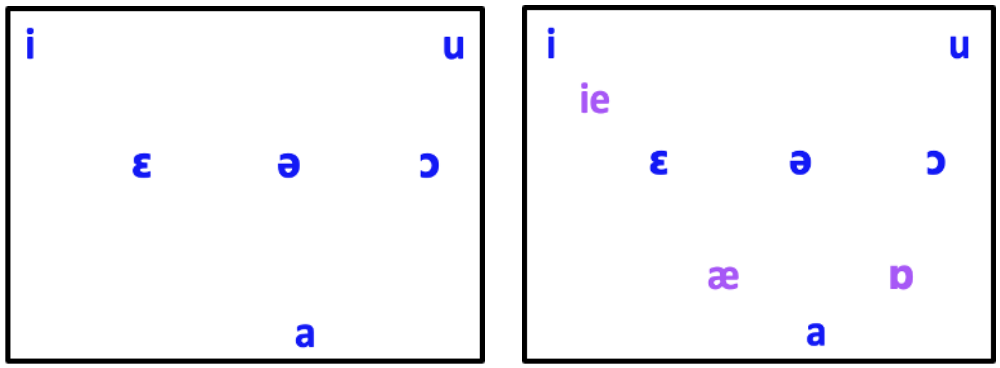


Figure 1.1: Vowel systems of SEA (left) (see Dum-Tragut 2009; Khach^{h}atryan 1988; Seyfarth et al. 2023) and the Gavaṛ (Bayazet) dialect (right) (based on descriptions of Katvalyan [2016]). Vowels not present in SEA are shown in purple. Not all illustrated distinctions are contrastive. The phonemic status of the schwa is debated (see Allen 1950; Dum-Tragut 2009; Khach^{h}atryan 1988; Khachatrian 1996; Vaux 1998). Some sources (e.g., Dum-Tragut 2009), transcribe the SEA mid vowels as [ɛ] and [ɔ], while Seyfarth et al. (2023: 19) describe the mid-front vowel as varying between [ɛ] and [e] and the mid-back vowel as acoustically closer to [o]. The transcription [ɔ] has been used for the sake of consistency with Portugal (2025).

As discussed in further detail in Chapters 2 and 3, this dissertation focuses primarily on the production and perception of the [ɒ] vowel in relation to both [a] and [ɔ]. The production studies examine the maintenance of the dialect vowel system as manifested through the maintenance of the vowel distinctions that differentiate it from SEA (including the distinctions between [a]/[ɒ]$_{A}$ and [ɔ]/[ɒ]$_{O}$), while the perception experiment explores how Gavaṛ residents perceive differences between the two members of each of these two vowel pairs. Given the differing phonological characteristics of these two vowel pairs (as described above), along with the social salience of the [ɒ]$_{A}$ allophone, investigation of these two vowel pairs allows for the exploration of the influence of phonological and social factors on production and perception. Gavaṛ residents produce these vowel pairs with a great deal of variation (based on both the 2022 and 2025 production results), such that some speakers may be moving toward a loss of the distinctions between the members of each pair. As such, the investigation of their production and perception also provides insights related to what it means for vowels to be distinct in a given language variety, with consideration for variation in perception and for social and phonological factors.

1.4 Overview of motivations from previous literature

1.4.1 Influence of social and phonological factors on production and perception

As noted above, the first overarching research question of this dissertation asks how social and phonological factors influence production and perception of vowels, as well as the relationship between production and perception. Social factors explored in this dissertation include demographic characteristics of speakers/listeners, social salience of linguistic features, and social information in the task, while phonological factors are explored via the investigation of different types of vowel differences (i.e., allophonic vs. contrastive).

It has long been established that speech perception is affected by social factors, such as listeners' knowledge/assumptions about a talker (e.g., Hay et al. 2006; Munson 2011; Schertz et al. 2019), or social salience of a linguistic feature (Juskan 2016). Speech perception has also been found to be affected by phonological information, as demonstrated by studies on perception of allophony (e.g., Boomershine et al. 2008; Peperkamp et al. 2003; Rose et al. 2023) and coarticulation (e.g., Beddor & Krakow 1999; Beddor et al. 2002; Kang et al. 2016). Social factors (see Eckert 2012) and language-specific phonology (e.g., language specific coarticulatory patterns – see Beddor et al. 2002) have also been found to affect speech production. However, despite this extensive previous literature, relatively few studies have examined the effects of social factors on perception of non-contrastive vowel differences (e.g., allophony and coarticulatory effects). This dissertation addresses this gap by investigating the perception and production of vowels with different phonological statuses, different relationships to coarticulatory processes, and different social characteristics.

1.4.2 Implications for understanding distinction between vowels and concepts such as "merger"

The second overarching question of this dissertation asks what it means for two vowel sounds to be distinct from one another in a given language variety, considering that both perception and production of vowels can be influenced by social and phonological factors. This includes questioning how perception data, including individual differences in perception, affects our understanding of what it means for two sounds to be distinct in production.

Importantly, this dissertation does not seek to define what it means for vowels to be contrastive (i.e., phonemically distinct). Rather, it seeks to contribute to discussions of how the size of acoustic differences between vowels with various phonological statuses and social characteristics relates to perception of those vowels, and of the implications of variation in perception (caused either by social and phonological factors, or by idiosyncrasies of individual

listeners) for how we interpret the significance of the size of acoustic differences. Though not all the vowel differences examined in this dissertation are contrastive, exploration of this question does have implications for how phenomena such as vowel "mergers" and "near-mergers" are understood.

There has been extensive previous study of vowel mergers, or situations in which two vowels that are contrastive become phonetically the same and lose the ability to distinguish word meanings (see Gordon 2013; Labov 1994: 324; Maguire et al. 2013). However, the criteria by which merger should be defined are debated – for example, previous authors have struggled to come up with non-arbitrary ways of defining how close two vowels must be acoustically for them to be considered "merged" (see Stanley & Sneller 2023). This lack of consensus about the implications of acoustic differences of a given size leads to a more general lack of conceptual precision in discussions of vowel mergers. For example, depending on different authors' preferred acoustic thresholds for considering two vowels "merged," one might deem a pair of vowels to be truly merged despite the existence of small but consistently measurable acoustic differences between them, while another author, in comparing such differences with listeners' failure to perceive them, might identify a case of near-merger (see further discussion in Chapter 3).

Literature on vowel mergers and near-mergers has generally not considered how the perception and production of non-contrastive vowel differences could inform such discussions – indeed, the investigation of the production and perception of such differences might initially seem like an unrelated issue. However, studying the perception and production of such differences is in fact relevant to debates about how to define merger, as even synchronically contrastive vowel differences can show remnants of phonetic conditioning (see discussion of the Bayazet dialect's vowel system above in Section 1.3.2.2 and discussion of Evanini's (2009: 141-142) remarks about the English low-back vowel merger in Section 3.4.1.1). Similarly, factors like coarticulatory influence of neighboring consonants could affect different vowel classes differently due to uneven distribution of phonetic environments in the lexicon. As such, consideration for such factors could have implications for the interpretation of small acoustic differences between vowels that appear to be "merged" in perception. In particular, if there are factors (such as unevenly distributed phonetic environments) that might prevent two vowels from ever becoming fully acoustically identical, how should we interpret the persistence of small

acoustic differences between them? Are these differences meaningful? In what cases is it justified to deem the vowels "near-merged" rather than just "merged"? More broadly, is it reasonable to expect that we can know and control for all other factors (e.g., social factors) that could lead to the maintenance of small acoustic differences between merging vowel classes? Similarities between the potential social functions of near-mergers and allophony have been noted by previous authors (Janson & Schulman 1983: 334; see also discussion in Chapter 3), and this dissertation seeks to build upon these observations by considering how the production and perception of allophony might inform definitions of merger and near-merger.

These questions are further complicated by previous findings that indicate the fluid nature of perception. As noted above, speech perception has been found to be influenced by both social and phonological factors. In the case of vowel mergers in particular, social information about the talker has been found to affect listeners' judgements of mergedness (e.g., Drager 2005; Hay et al. 2006; Koops et al. 2008). Thus, determining whether vowels are merged in perception is hardly simpler than determining if they are merged in production, and consideration for social factors, as well as factors such as task type and individual variation in perception, introduces a great deal of uncertainty. Accordingly, if perceptibility of differences between vowels is considered crucial to the definition of merger (see Gordon 2013: 205), then what are the implications of variation in perception, both across and within individuals? How does the variability of perception affect our interpretation of the meaningfulness of small acoustic differences in production? Just like differing and arbitrary definitions of the size of acoustic difference that equates to merger, differences in methodology for determining that listeners have a perceptual merger could lead to differing conclusions about the status of a given vowel phenomenon. This dissertation thus attempts to consider the complexities of variation in both perception and production in commenting on what it means for vowels to be distinct in a given language variety.

1.5 Outline of dissertation structure

This dissertation consists of seven chapters. Chapter 2 provides an overview of the results of the 2022 production study (which are presented in more detail in Portugal [2025]), before presenting the results of the production data collected alongside the perception experiment in 2025. In comparing the results of the two studies, the chapter discusses patterns of social variation in the maintenance of dialect vowels found in the two data sets, while also reflecting on more general questions related to sociolinguistic research (e.g., implications of participant

sampling methods, recording quality, elicitation methods) that emerge from the comparison of the results.

Chapter 3 introduces the specific questions explored in the perception experiment, which are as follows:

1) How do residents of Gavaṛ discriminate differences between various vowels of the local dialect's vowel system?

2) How do residents of Gavaṛ categorize such vowels in a task that explicitly elicits metalinguistic judgements about features of the local dialect?

3) How might participants' discrimination of vowel differences in 1) relate to or differ from their categorization of vowels in 2)?

4) How do participants' productions of vowel differences compare with their perceptions, and how might individual participants differ in this regard?

The chapter motivates these questions through discussion of the 2022 production results and through a review of relevant literature (on topics such as vowel mergers and near-mergers, links between production and perception, perception of allophony and coarticulation, and the role of social information in speech perception). This chapter also presents a broad overview of the experiment's methods and procedures, as well as a broad overview of the data collection process.

Chapter 4 presents the goals, expectations, methods, and results for the AX discrimination task. The results presented in this chapter come from quantitative analyses of responses to this task. In particular, the methodology included the calculation of various metrics of performance on the task, which were then used as outcome variables in statistical analyses, as well as statistical methods predicting raw responses to the task (mixed-effects linear and logistic regression models). The quantitative analyses presented in this chapter also include some exploration of relationships between production metrics and performance on the task.

Chapter 5 presents the goals, expectations, methods, and results for the dialect categorization task. The results presented in this chapter come from statistical analyses predicting responses to the given task. Specifically, the chapter presents both group-level analyses predicting raw responses to the task (mixed-effects logistic regression models), as well

as additional analyses in which individual-level metrics of performance on the task are used as outcome variables in statistical models. The statistical models presented in this chapter also include exploration of relationships between production metrics and performance on the task, as well as exploration of whether performance on the AX task predicted performance on the dialect categorization task.

Chapter 6 investigates how perception compares across the two experimental tasks and how production relates to perception. First, the chapter summarizes the main findings of Chapters 4 and 5, in order to draw broad comparisons related to how perception of the relevant vowels differed across the two tasks. The chapter also summarizes the results of statistical analyses from Chapters 4 and 5 that relate either to production-perception relationships or to relationships between performance on the two tasks. Demographic factors that were found to be relevant for explaining variation in production (presented in Chapter 2) are also compared with those found to be relevant for predicting responses to the perception tasks in Chapters 4 and 5. Finally, this chapter presents the results of two additional qualitative individual-level analyses. The first of these explores how perception differed across the two tasks by comparing each participant's responses to the same stimuli in each task. The second explores production-perception relationships by comparing participants' productions with their own performance on the AX task and with the performance of the most accurate discriminators in the sample, while also comparing participants' perception and production with criteria for discriminability and/or mergedness used in previous literature. In connection with the results of this analysis, implications of individual and community-level production-perception relationships for defining vowel distinction and merger are discussed.

Finally, Chapter 7 summarizes the results of all analyses and discusses how the results relate to the broad questions of the dissertation, expounding upon the broader implications and contributions of this work.

Chapter 2 Maintenance of Previously Described Dialect Vowels in Production: Comparison of Demographic Patterns in 2022 and 2025 Data

2.1 Chapter summary

This chapter provides an overview of two production studies, which elicited Gavaṛ residents' pronunciations of various vowels with the goal of exploring social variation and assessing maintenance of the dialect vowel system described by Katvalyan (2016). The chapter identifies patterns of variation in the use/maintenance of dialect vowels in both data sets (collected in 2022 and 2025), while also commenting on methodological insights that can be gained from comparing the results of the two studies with consideration for their differing methodologies. The chapter begins with a brief overview of the goals of both production studies, after which a summary of the results of the 2022 sociolinguistic interviews is presented. The methods and results of the 2025 production study (recorded word list) are then presented, followed by comparisons of the 2022 and 2025 studies that discuss demographic patterns found in relation to various vowels, as well as lessons about elicitation techniques, token sample sizes, participant sampling, and recording quality that can be gleaned from the comparison of the two data sets.

Though the 2025 analysis only included a subset of the vowels included in the 2022 analysis, the results of the two analyses were similar overall. Gender-based variation (with men using more dialect-like vowels than women) was found for the vowel pairs [a]/[ɒ]$_{A}$ and [ɒ]$_{A}$/[ɒ]$_{O}$ in both studies, while such variation was found for [ɔ]/[ɒ]$_{O}$ only in the 2022 data (see Chapter 1 and Section 2.3.2.1 below for overview of the Gavaṛ dialect's vowel system). This discrepancy is likely explained by the participant sampling issues discussed below (Section 2.6.1). Other discrepancies between the two data sets (e.g., differing overall magnitude of Pillai scores) are also discussed in relation to the potential influence of factors such as token sample size and recording quality. More general implications of the differing results of the two studies are also discussed, as the discrepancies suggest that subtle aspects of elicitation method (e.g., the broader context in which a word list is recorded), as well as factors like participant sampling method,

token sample size, and recording equipment can have substantial effects on the results of acoustic analyses.

2.2 Goals and research questions of production studies

The 2022 production study sought to evaluate maintenance of the dialect vowel system as described by Katvalyan (2016). Two questions were explored:

1) Which demographic factors (among GENDER, BIRTH YEAR, and EDUCATION LEVEL) relate to maintenance of the previously described dialect vowel system?

2) How does participants' metalinguistic commentary inform our understanding of observed correspondences between demographic factors and dialect maintenance?

This chapter primarily discusses the methods and results related to the first research question, as they are necessary for interpreting the motivations and results of the 2025 production study. Discussion of the second question (as well as a more complete overview of the methods and results of the 2022 production study) can be found in Portugal (2025).

As discussed below, analysis of the 2022 data (specifically, of picture-naming and word-list data) showed that some speakers maintained some of the dialect vowels described by Katvalyan (2016). Demographic patterns were found in relation to GENDER for some vowels, with men maintaining dialect vowels more than women. The sections below also discuss variation in the vowel [æ], the analysis of which was not presented in Portugal (2025). Variation in this vowel, which appears to be undergoing a "merger by transfer," was found to relate to both GENDER and EDUCATION LEVEL (see Section 2.3.3.2). A preliminary analysis of one dialect vowel in spontaneous speech data found both gender- and age-based variation (with older people using more dialect-like pronunciations), suggesting that style-shifting may have prevented age-based variation from being captured in the main analysis (see discussion in Section 2.3.3.1).

The goals of the 2025 production study were twofold:

1) To further explore relationships between dialect vowel maintenance and demographic factors, comparing demographic patterns found in the 2022 and 2025 data

2) To compare participants' own productions of certain vowels with their perception of those vowels, as explored in the 2-part perception experiment

As discussed below, the 2025 production study focused on analysis of those vowels that were used as test stimuli in the perception experiment, rather than encompassing analysis of all dialect vowels. The analyzed vowels were also those for which gender-based variation was found in the main analysis of the 2022 data. Though gender-based variation was also found in relation to [æ] in the 2022 data, this vowel was used for training and filler stimuli in the perception experiment for reasons discussed in Chapters 4 and 5 (thus, it is not included in the main analysis of the 2025 data, though participants' productions of [æ] in 2025 vs. 2022 will be discussed in Section 2.6.3 in relation to what they tell us about the implications of different elicitation techniques).

The remaining sections of this chapter present the methods and results of the 2025 production study in comparison with the relevant methods and results of the 2022 study. The overarching goal of the chapter is to describe patterns of social variation in the maintenance of dialect vowels in Gavaṛ based on the available data sets, with a particular focus on those vowels further explored in the perception experiment. This analysis thus sets the stage for comparisons of production and perception that are investigated in later chapters. Via comparison of two data sets that had partially overlapping methodologies and participant samples, this chapter also contributes to conversations about methodological issues in sociolinguistic research. Section 2.3 briefly discusses the data collection methods of the 2022 study, followed by a more in-depth presentation of analysis methods that were used in both the 2022 and 2025 studies, and a very brief discussion of the 2022 results. This section also presents the analysis and results related to [æ] in the 2022 data, as these were not presented in Portugal (2025). Sections 2.4 and 2.5 present the methods and results of the 2025 analysis, while Section 2.6 discusses these results in comparison with the 2022 results, as well as the methodological implications of this comparison. Comparison with the results of the perception experiment is presented in Chapters 4-6.

2.3 Summary of 2022 methods and results

2.3.1 Data collection

The 2022 study consisted of 31 sociolinguistic interviews conducted by a local research assistant. As speakers are known to accommodate to their interlocutors (Bell 1984), interacting with an interviewer who spoke a different variety of Armenian would likely have caused

participants to alter their speech.[9] All participants were recruited from the interviewer's existing social networks. The interview, whose structure was adapted from the classic sociolinguistic interview (see Becker 2017) had four parts: 1) conversation portion, 2) picture-naming task, 3) word list, and 4) metalinguistic questions. The word list elicited phonetic features, while the picture task elicited phonetic features and lexical variation. Though sociolinguistic interviews traditionally make use of a reading passage and sentence list (in which target words are embedded into a carrier sentence), participants in this study were instead asked to name a series of pictures and to read a list of words presented in isolation. This method was chosen in order to minimize the influence of standard Armenian grammar/vocabulary/orthography on participants' pronunciations. Words in both lists were attested by Katvalyan (2016, 2009) or proposed/verified by local research assistants.[10] Minimal and near-minimal pairs targeting vowels were elicited, as well as additional words containing each vowel.

Participants signed an informed consent form before the interview, and after the interview completed an additional consent form where they could give permission for various specific uses of their data. The elicitation materials and consent forms were reviewed by the University of Michigan IRB (HUM00214786), and the study was determined to be exempt from ongoing IRB review.

2.3.2 Analysis of maintenance of vowel distinctions

2.3.2.1 Conceptual basis

Both the 2022 and 2025 analyses examined each participant's adherence to the dialect vowel system by conceptualizing the differences between the dialect and SEA systems as a set of phonemic and allophonic distinctions that are present in the dialect but not in SEA. In the first step of the analysis, each speaker's data was coded according to the vowel expected in each word based on previous dialect descriptions (Katvalyan 2016, 2009), rather than in accordance with the speaker's actual production. For example, if the vowel [ɒ] was expected in a given word based on the previous descriptions, the vowel was coded as such, even if the speaker produced a

[9] Having learned Armenian as an adult and speaking a mixture of SEA and the colloquial variety of Yerevan, I would not have been a suitable interviewer given this goal of avoiding interlocutor accommodation.
[10] Katvalyan's (2021) new dictionary of the Bayazet dialect was not available at the time of the fieldwork, despite the publication date being listed as 2021. The dictionary was retroactively used to determine the predicted pronunciation of one dialect variant (*dæɾgæχ*, 'gate'), which was used by many participants despite not having been expected based on Katvalyan (2016, 2009).

vowel closer to [a]. Coding tokens in this way allowed for the mapping of each speaker's actual vowel space onto the predicted dialect vowel space.

Based on the largely overlapping lexicon of SEA and the dialect, it is possible to say which SEA vowel each dialect vowel is likely to be moving toward in speakers who are shifting toward SEA. Thus, in order to assess each participant's maintenance of dialect vowels, words predicted to contain the additional vowels of the dialect were compared with words containing the corresponding SEA vowels. Four such comparisons were made in the analysis of the 2022 data:

1) Words where [ie] was predicted in the dialect were compared with words where [ɛ] was predicted in both varieties

2) Words where [ɒ] was predicted in the dialect while [a] appears in SEA (labeled as [ɒ]$_A$ throughout this dissertation) were compared with words where [a] is predicted in both varieties

3) Words where [ɒ] was predicted in the dialect while [ɔ] appears in SEA (labeled as [ɒ]$_O$ throughout this dissertation) were compared with words where [ɔ] is predicted in both varieties

4) The words where [ɒ]$_A$ and [ɒ]$_O$ were predicted were compared with one another, as according to Katvalyan's (2016) description, these two groups of words should have an identical vowel

Thus, there were three comparisons ([ie]/[ɛ], [a]/[ɒ]$_A$, and [ɔ]/[ɒ]$_O$) in which more distinction corresponds to more adherence to the dialect vowel system, and one comparison ([ɒ]$_A$/[ɒ]$_O$) in which less distinction equates to more adherence to the dialect vowel system. These comparisons are conceptualized in **Figure 2.1**.

In the analysis of the 2022 data, it was originally planned that words where [æ] was predicted in the dialect would also be compared with words where [a] was predicted in both varieties. However, the words where [æ] was predicted could not be analyzed in the same manner as the words where other dialect vowels were predicted due to bimodality of this group of tokens for many speakers. Specifically, in the 2022 data, around two thirds of the participants produced two distinct groups of pronunciations among these tokens (corresponding to [æ] and [a]), thus exhibiting characteristics of an ongoing "merger by transfer" (see Section 2.3.2.3). Because Katvalyan (2016: 83) describes [æ] as a variant of [a] despite its phonemic status, and

because [æ] corresponds to SEA [a] in cases where vocabulary is shared across the two varieties, it was indeed expected that words where [æ] was predicted in the dialect would contain [a] in cases of convergence toward SEA. However, the presence of clusters of categorically different pronunciations among the group of tokens where [æ] was predicted meant that the comparison of the [æ] and [a] classes could not be analyzed in the same manner as the other comparisons. The analysis of this pair is thus discussed separately in Section 2.3.2.3.

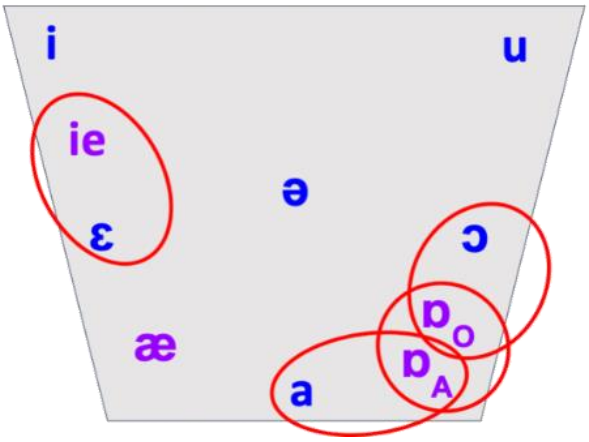


Figure 2.1: Pairs of vowels examined in the 2022 analysis. The vowel system of SEA is shown in blue (see Dum-Tragut 2009; Khach^{h}atryan 1988; Seyfarth et al. 2023), while the additional vowels of Gavaṛ's dialect are shown in purple.

In light of the heterogeneous language situation in the Republic of Armenia, in which both SEA and the colloquial variety of Yerevan could potentially be the target of those speakers who are shifting away from local dialects, it is important to note that the present analysis accounts at least to some extent for the possibility of convergence toward the vowel system of either variety. This is because the contemporary variety of Yerevan, as well as many varieties of the Ararat dialect group (which is spoken around Yerevan, and on which SEA is based) have the same six-vowel system as SEA (see Seyfarth et al. [2023] for discussion of contemporary Yerevan Armenian and Markosyan [1989: 53] for discussion of the Ararat dialect group). Based on Hodgson's (2019) observation about potential ongoing dialect leveling in regional cities, it was expected that some convergence toward this six-vowel system would be observed. In order to quantify this convergence and compare it across speakers, the quantitative analysis described in the following section was implemented.

2.3.2.2 Quantitative comparison of speakers

2.3.2.2.1 Calculating metrics of distinction/merger

Before quantitative comparison of the maintenance of dialect vowels across speakers could be undertaken, data processing and measurement of the first and second formant frequencies were carried out according to the procedure described in Portugal (2025) (see also Section 2.4.3.1 below for detailed discussion of the very similar procedures used in processing the 2025 data). The analysis included only the picture and word list tokens (see Portugal [2025] for explanation of the exact criteria determining which tokens were included in the analysis, and see Section 2.3.3.1 and Portugal [2025] for discussion of a small preliminary analysis of spontaneous speech from the conversation portions of the interviews).

Analysis of the four comparisons depicted in **Figure 2.1** was operationalized through the calculation of Pillai scores (to measure overlap between clusters) and Euclidean distances (to measure distances between cluster centers) between each vowel pair for each speaker. As Nycz and Hall-Lew (2014) explain, measuring both overlap and distance is advisable because they can vary independently of each other. These static measures were implemented despite Katvalyan's (2016) description of several of the relevant vowel classes ([ie] and [ɒ]) as diphthongal due to the relative infrequency of obviously diphthongized pronunciations in the analyzed tokens (see Chapter 1 for discussion of Katvalyan's description of these vowels, and see Section 2.3.3.1 below for discussion of audibly diphthongal pronunciations in the 2022 data).

Pillai scores (formally known as the Pillai-Bartlett trace) were first used to measure vowel overlap by Hay et al. (2006) and have become a "standard metric" for this purpose (Stanley & Sneller 2023: 54), having been found by Kelley and Tucker (2020: 143) to be the preferred method especially when sample sizes are small. The Pillai-Bartlett trace is a test statistic of multivariate analysis of variance (MANOVA) models, which test the significance of category membership in explaining the distributions of data (see Nycz & Hall-Lew 2014; Stanley & Sneller 2023). Pillai scores range between 0 and 1, with 0 indicating full overlap and 1 indicating complete lack of overlap. Euclidean distance has also commonly been used to study vowel mergers and distinctions (Nycz & Hall-Lew 2014).

Both Pillai scores and Euclidean distance can be calculated in ways that account for the potential effects of neighboring consonants on vowel formants. During the calculation of Pillai scores, phonological context can be included as a predictor in MANOVA, which allows variation

attributable to this context to be factored out of the Pillai score calculated in relation to the category (i.e. word class, vowel class) predictor (Nycz & Hall-Lew 2014: 5). The "adjusted Euclidean distance" method developed by Nycz (2011) and further described by Nycz and Hall-Lew (2014) has a similar goal. This method uses mixed-effects linear regression models with either F1 or F2 as the dependent variable. Nycz and Hall-Lew (2014) propose comparing a model that contains fixed effects related to phonological environment and a random effect for "word" with a model that includes an additional fixed effect for "word class." As this "word class" fixed effect represents the two classes of words under comparison, its effect size in the second model represents the adjusted distance between the two classes for whichever formant is the dependent variable (Nycz & Hall-Lew 2014).

Although these methods have clear advantages in comparison with the calculation of simple Pillai scores and Euclidean distance, their implementation proved problematic in the present study. It would have been ideal to be able to control for preceding and following place of articulation in the analysis of the 2022 production data in particular (compared with the 2025 production data, which comes only from a word list and is thus more uniform across participants). Though the picture and word list tokens are much more similar across speakers than they would be in the case of using, for example, spontaneous speech data, participants still had different sets of tokens due to variation in the naming of pictures. However, two factors prevented the implementation of methods that would control for phonetic context.

First, these methods are most appropriate for the analysis of phonemic vowel distinctions, so that the variation in the data accounted for by phonetic context and the variation accounted for by word class will not be expected to overlap. As many of the comparisons explored in the present analyses are allophonic or show remnants of historical allophony, the preceding and/or following places of articulation are part of what define the vowel classes, meaning that phonetic context and vowel class cannot be separated in the analysis. For example, the comparison of [a]/[ɒ]$_{A}$ is completely allophonic. Although most of the allophony is influenced by features such as aspiration, voicing, and nasality, the [ɒ]$_{A}$ allophone also occurs in the word-initial position, meaning that the absence of a preceding place of articulation is associated only with the word class where [ɒ]$_{A}$ is expected. Thus, if a predictor encoding preceding place of articulation included a level for "zero" place of articulation, this predictor would account for variation that should actually be attributed to the word class. A similar problem is presented by the comparison

of [ɔ]/[ɒ]o. Although this distinction is phonemic according to Katvalyan (2016), a subset of relevant words behaves in a way that is essentially allophonic. In monosyllabic words that are not recent borrowings, [ɒ]o is expected unless the following consonant is [v] (the only labiodental consonant found in this position in the words represented in the present study), in which case [ɔ] is expected. Thus, if a predictor encoding following place of articulation included a "labiodental" level, then this predictor might account for variation that is actually attributable to word class.

Second, the per-speaker sample sizes representing some preceding and following places of articulation in the present data sets are simply very small (in some cases, only 1-3 tokens per speaker per vowel class). The sample sizes of different places of articulation are also in some cases very unbalanced across the vowel classes being compared (even in cases other than those of full and partial allophony described above). These issues would call into question the reliability of an analysis that attempted to incorporate the influence of place of articulation.

Accordingly, the present analysis employed simple Pillai score and Euclidean distance calculations. An additional open question in sociophonetic research is whether or not formant values (generally measured in Hz) should be transformed to the Bark scale before being subjected to further analyses. The Bark scale is a perceptual scale that is commonly used in the study of formants, as it transforms Hz values in a way that approximates the human auditory system's non-linear response to them (see Thomas 2011: 56-58). While some studies (e.g., Hay et al. 2006) convert Hz values to Bark before carrying out additional analyses (in the case of Hay et al. [2006] – the calculation of Pillai scores), others do not. For example, Kendall and Fridland (2017) calculate Euclidean distance using Lobanov-normalized formant measurements (which according to the available information were not converted to Bark prior to normalization).

Though it is not one of the main goals of this dissertation to argue for or against the conversion of formants to the Bark scale, the potential effects of such conversion are explored for the sake of methodological and analytical thoroughness. As such, unless otherwise noted, all Pillai score and Euclidean distance calculations were carried out both with and without conversion of Hz values to Bark. Conversion of Hz values to Bark was carried out using the *barktools* package (Stanley 2020). In the case of Pillai scores, it was found that conversion to Bark does not affect the calculated values (beyond very small differences attributable to rounding during the Bark conversion). As the calculation of Euclidean distance requires

normalized formant measurements, conversion to Bark was applied before the Nearey1 normalization procedure (described in Portugal [2025] and in Section 2.4.3.1 below). Though a normalization procedure specifically meant to be implemented with Bark-converted formants exists (the Bark Difference Metric – see Thomas 2011: 163-165), this procedure requires F3, which was not consistently measurable in the present data sets due to varying and sometimes suboptimal recording quality. Nevertheless, one can also find examples of Bark-converted values being subjected to other normalization procedures – for example, Hall-Lew (2010) converts formant values to Bark before subjecting them to Lobanov normalization. In the case of the 2022 production data, it was found that conversion to Bark did not affect the outcome of the analysis (in terms of which social factors were found to relate to the use of more dialect-like vowels). However, in the case of the 2025 production data, both versions of the analysis will be discussed, as the conversion to Bark before calculation of Euclidean distance did sometimes affect the results.

Pillai scores were calculated in R (R Core Team 2021) for each speaker for each vowel comparison. Raw (unnormalized) data was used for the calculation of Pillai scores, as normalization does not affect the resulting scores.[11] Visualizations of Pillai scores presented in Portugal (2025) and throughout this dissertation also include individual merger/distinction thresholds calculated using the method proposed by Stanley and Sneller (2023), which accounts for sample size. According to Stanley and Sneller, this method addresses the decreased likelihood of finding low Pillai scores (i.e. very overlapping categories) when sample sizes are small. This method establishes the Pillai score below which the score of samples of a given size from a truly merged underlying population will fall in 95% of cases, and is calculated as $\frac{e}{m}$, where *m* is the average sample size per group. Although it is also intended to address the lack of any widely accepted, non-arbitrary Pillai score merger/distinction threshold, this method also ultimately seems arbitrary, as it nevertheless still relies on an arbitrary statistical significance threshold. It is also not clear if vowels deemed distinct by this method would always be perceived as distinct (see further discussion in Chapter 6). These thresholds thus lack crucial information for interpreting sociolinguistic data, and they are included in graphs below as a tool

[11] I tested this myself, but see also Stanley (2021).

to visualize the effects of sample size rather than as a definitive means of deeming speakers to have merged vs. distinct vowel categories.

Euclidean distance measurements were calculated individually using the normalized data for each speaker for each vowel comparison. Following Nycz and Hall-Lew (2014), Euclidean distances have been calculated using median formant values, which are less skewed than means by extreme values and non-normal distributions. The calculated Euclidean distance values are thus referred to as ED-median throughout the remainder of this dissertation.

2.3.2.2.2 Comparison of speakers with linear regression models

The three demographic factors included in both the 2022 and 2025 analyses – GENDER, BIRTH YEAR, and EDUCATION LEVEL – represent key macro-social factors that have often been included in sociophonetic research (see Kendall et al. 2023). In the 2022 study, GENDER and EDUCATION LEVEL were operationalized as categorical variables with two and three levels respectively (men vs. women and secondary education vs. vocational education vs. higher education), while age (BIRTH YEAR) was a continuous variable. In the 2025 study, EDUCATION LEVEL only had two levels (higher education and non-higher education), for reasons discussed in Section 2.4.1 below. Information related to all three variables was self-reported by participants. EDUCATION LEVEL was used in place of the more commonly used categories of social class or socioeconomic status in an attempt to capture the realities of post-Soviet Armenia. Unemployment and underemployment remain serious issues as a result of post-Soviet economic collapse, such that many people (including many participants of the present study) do not work in the fields in which they were educated. Though Labov (2001: 60) notes that occupation is generally accepted to be "the most highly correlated with other conceptions of social class," previous work in other post-Soviet states also calls into question the appropriateness of assigning people to social classes based on characteristics such as their occupation or income. For example, Garibova (2020: 65) notes a distinction between social prestige and socioeconomic status in post-Soviet Azerbaijan, identifying its roots in "the Soviet separation of economic status from socio-cultural identification based on parameters such as education, erudition, language skills, and ability to understand and value art."

Linear regression models were used as a means of quantitatively comparing speakers' maintenance of dialect vowels. Both Pillai scores (e.g., Baranowski 2022; Berry & Ernestus 2018; D'Onofrio et al. 2019; Helms 2023; Labov et al. 2016; Lee 2016; Wong & Hall-Lew 2014;

Xu 2019) and Euclidean distances (e.g. Baranowski 2022; Lee 2016; Xu 2019) have been used as the dependent variables in linear models in many previous sociophonetic studies.

Two fixed-effects linear regression models were created for each vowel comparison – one model with Pillai scores as the dependent variable and one with ED-median as the dependent variable. The statistical significance of the predictors in these models provides evidence for which demographic groups can be said to be more closely adhering to the dialect vowel system. For example, if the predictor of GENDER is found to be statistically significant in a given model, it will tell us that one of the two gender groups is adhering more closely to the dialect vowel system, while the lack of a significant result will suggest that neither of the gender groups is adhering to the dialect vowel system more than the other. The regression coefficients will tell us which group is adhering more closely to the dialect vowel system, as these coefficients will indicate which of the two groups had larger vs. smaller Pillai scores or ED-median values. For the three comparisons of [ie]/[ɛ], [a]/[ɒ]$_A$, and [ɔ]/[ɒ]$_O$, higher Pillai scores and ED-median values, which correspond to less overlap and more distance between the vowel categories, indicate more adherence to the dialect vowel system, while the opposite is true for the comparison of [ɒ]$_A$ and [ɒ]$_O$.

All models were implemented in R. In the comparison of education groups in the 2022 data, higher education served as the reference level against which the vocational and secondary groups were compared. The *contrast()* function from the *contrast* package (O'Callaghan et al. 2022) was applied to the model outputs in order to test the difference between the two non-reference groups. Multicollinearity of predictors was tested using the *vif()* (variance inflation factor) function from the *car* package (Fox & Weisberg 2019). No interactions were included in the analysis of the 2022 data, both because of the exploratory nature of the analysis and lack of specific hypotheses related to any interactions, and because of the very small number of participants possessing some combinations of characteristics (for example, the sample included only two women with secondary education). Interactions and additional predictors were tested in the 2025 analysis, as discussed below in Section 2.4.3.1.

2.3.2.3 Analysis of "merger by transfer" of [æ] and [a] in 2022 data

As noted above, the group of tokens in which [æ] was predicted could not be included in the main analysis of the 2022 data. This was because many speakers pronounced this group of tokens in a clearly bimodal manner, sometimes pronouncing sounds closer to [æ] and sometimes

pronouncing sounds closer to [a]. The widespread bimodality was verified via examination of three-dimensional density plots that were produced in R with the *rayshader* package (Morgan-Wall 2020) and *ggplot2* (Wickham 2016). These density plots visualized the group of tokens where [æ] was expected, and for comparative purposes included an overlaid visualization of each participant's group of tokens where [a] was expected (see example below in **Figures 2.2-2.4**). Based on an examination of these plots, it was determined that 21 out of 31 participants exhibited some degree of bimodality, in which one portion of the tokens where [æ] was expected overlapped with those where [a] was expected, and one portion did not overlap, being located in an area of the vowel space more consistent with an [æ]-like pronunciation.[12]

The presence of distinct clusters of [æ]-like and [a]-like pronunciations among this group of tokens for some speakers suggests that [æ] and [a] are undergoing a "merger by transfer" in Gavaṛ. Labov (1994: 321) defines merger by transfer as a process in which "words are transferred gradually from one phonemic category to another," without intermediate phonetic forms. Labov notes that such vowel mergers generally constitute a "change from above" (i.e. above the level of consciousness), as they often involve linguistic variants that have "acquired a social stigma or prestige" (1994: 321). Maguire (2007: 325) notes that merger by transfer often occurs in situations where speakers of a dialect are shifting their phonological system towards a "standard or supralocal variety," while Herold (1997: 186) states that dialect contact is a "necessary condition" for merger by transfer. The other two types of vowel mergers identified in previous literature are "merger by drift (or approximation)" and "merger by expansion" (see Maguire et al. 2013: 232). In merger by drift/approximation, the phonetic spaces of the relevant vowel categories gradually move closer to one another; the process is thus "phonetically gradual" (i.e., involving intermediate phonetic forms), but, unlike merger by transfer, is not "lexically gradual" (Maguire et al. 2013: 232-233). In merger by expansion (first identified by Herold [1990]; see also Maguire et al. [2013: 233]), the phonological distinction between two vowels suddenly collapses, and the entire phonetic space of the two original phonemes becomes associated with the new merged phoneme. Though the term "merger" is generally used to discuss cases (such as the merger of [æ] and [a] in Gavaṛ) involving the loss of phonemic contrast, a

[12] No speaker who exhibited a unimodal distribution of the tokens where [æ] was expected produced these tokens in a separate area of the vowel space from the tokens where [a] was expected. In other words, in the given group of participants, unimodal distributions are equivalent to only using an [a]-like pronunciation for both the [æ] and [a] groups of tokens.

review of previous literature on mergers by approximation also raises questions related to variation in the use of non-phonemic vowel differences (such as the difference between [a] and [ɒ]$_A$). As such, literature on mergers by approximation and "near-mergers" is further discussed in Chapter 3.

Although the three types of vowel merger are often discussed as separate processes, it is important to note that the reality of mergers in progress is often more complex. For example, although many participants of the 2022 study showed a distribution of [æ] tokens that seemed clearly consistent with an ongoing merger by transfer (e.g., **Figure 2.2**), others showed somewhat bimodal distributions that nevertheless seemed more consistent with a merger by approximation due to the apparent presence of intermediate phonetic forms, which could be indicative of a phonetically gradual merger (see **Figure 2.3**). Such mixture of mergers by transfer and approximation among a single community has been observed by other authors (e.g., Britain & Grossenbacher 2021).

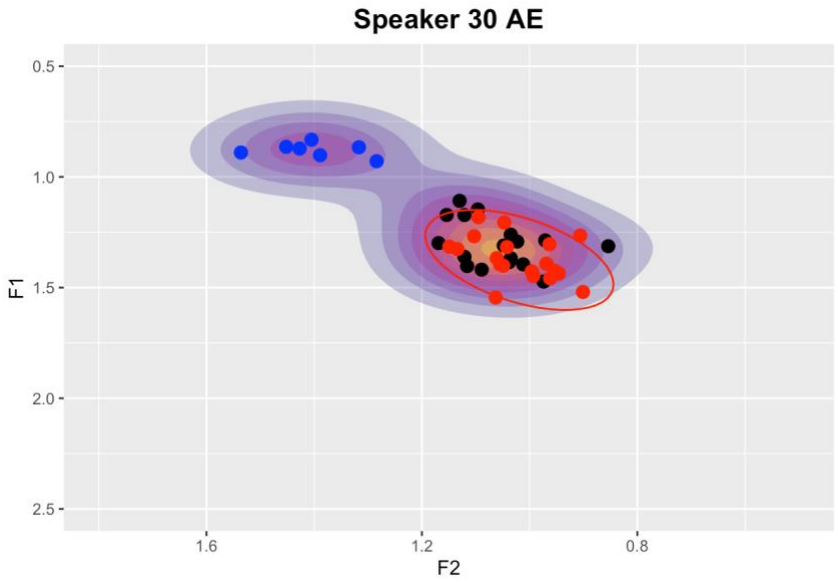


Figure 2.2: The [æ] and [a] tokens of a participant exhibiting a merger by transfer of [æ] and [a]. The tokens where [a] was predicted are represented by red dots and a red ellipse (representing 2 standard deviations of this token category), while the blue and black dots represent the tokens where [æ] was predicted. The black dots represent those [æ] tokens that were clustered together with the [a] tokens by k-means analysis. The 3-dimensional density plot represents the distribution of only those tokens where [æ] was predicted.

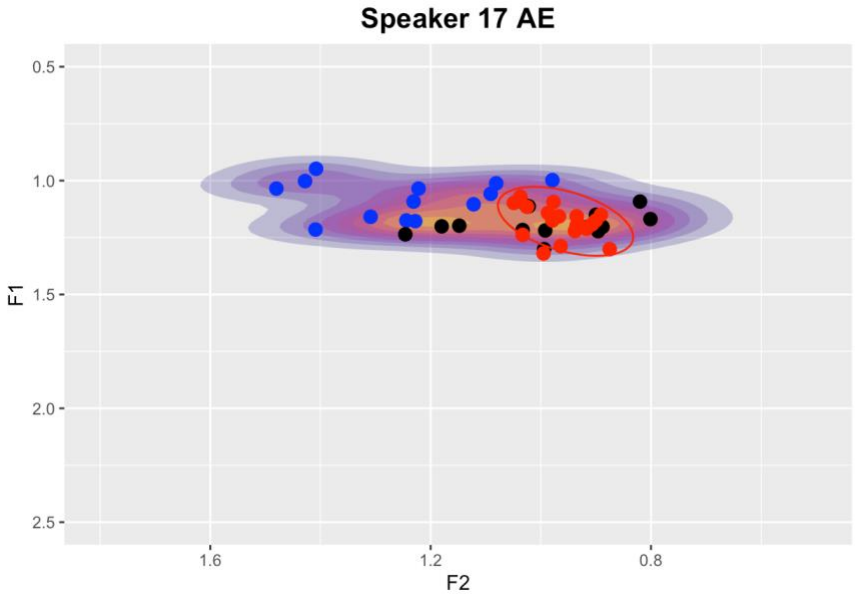


Figure 2.3: The [æ] and [a] tokens of a participant exhibiting a merger by approximation of [æ] and [a]. The tokens where [a] was predicted are represented by red dots and a red ellipse (representing 2 standard deviations of this token category), while the blue and black dots represent the tokens where [æ] was predicted. The black dots represent those [æ] tokens that were clustered together with the [a] tokens by k-means analysis. The 3-dimensional density plot represents the distribution of only those tokens where [æ] was predicted.

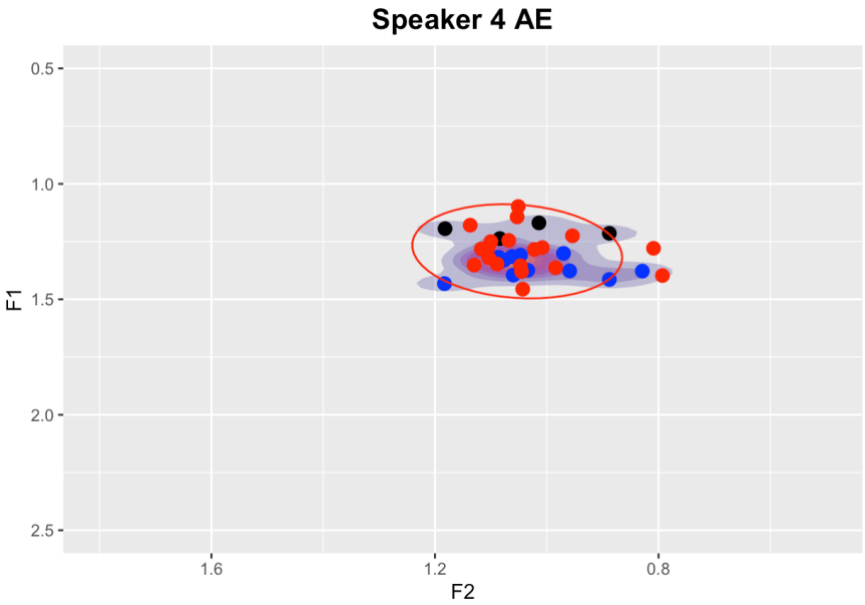


Figure 2.4: The [æ] and [a] tokens of a participant exhibiting a complete merger of [æ] and [a]. The tokens where [a] was predicted are represented by red dots and a red ellipse (representing 2 standard deviations of this token category), while the blue and black dots represent the tokens where [æ] was predicted. The 3-dimensional density plot represents the distribution of only those tokens where [æ] was predicted.

Given the vast variation among participants exemplified by **Figures 2.2-2.4**, a method was developed in order to examine the degree to which different speakers could be described as using distinct [æ] and [a] sounds, which would indicate adherence to the previously described dialect vowel system. The overarching idea of this analysis was to group each speaker's relevant tokens into two clusters, the distance between which could then be measured. The resulting distances could then serve as a metric of the degree to which a speaker is producing distinct [æ] and [a] tokens; the larger the distance, the more justified we are in supposing that the two clusters represent distinct sounds, and thus in supposing that the speaker is using a vowel system closer to the one described by Katvalyan (2016).

The data were first normalized with Nearey1 normalization (for additional discussion of formant measurement and normalization, see Portugal [2025] and Section 2.4.3.1 below).[13] Next, each participant's individual F1 and F2 data sets including the tokens in which either [æ] or [a] was predicted were z-scored, so that the two axes would be on an identical scale. Each participant's data was then subjected to k-means analysis to group the tokens into two clusters. It was necessary to include the tokens where [a] was predicted in this analysis (as opposed to including only those tokens where [æ] was predicted) because the set of tokens where [æ] was predicted was relatively small. Some participants who exhibited a bimodal distribution of the tokens where [æ] was predicted thus had very small numbers of tokens (e.g., 1-3 tokens) that visually corresponded to either [æ] or [a] pronunciations. Such small sample sizes would have created difficulties for the clustering analysis, including making the locations of identified cluster centers unreliable in cases where clusters consisted of very few tokens. Despite the pooling of [æ] and [a] tokens in the analysis, one participant (Speaker 22) nevertheless needed to be excluded, due to having only two tokens corresponding to an [æ] pronunciation (from which a cluster center location would not have been able to be reliably calculated).

Z-scoring was necessary in order for the clusters identified by k-means to correspond to visual clusters in the data. This is because k-means analysis groups tokens by minimizing the

[13] Two words where [æ] was initially predicted, but which ultimately could not be clearly assigned to a token category, were excluded from the data set after normalization. The word *kvas* ('kvas, pickled vegetables'; possible pronunciations: [kəvas], [kəvæs]) as elicited by both the picture and word lists was omitted due to conflicting information from Katvalyan (2016, 2021) and local research assistants about the predicted dialect pronunciation. The word *t^{h}at^h* ('paw'; possible pronunciations: [t^{h}at^h], [t^{h}æt^h]) has been omitted in cases where it was elicited by the word list, as the orthography is ambiguous and could indicate either the intended word (predicted dialect pronunciation: [t^{h}æt^h]) or a dialectal word meaning "person who does not understand" with the expected pronunciation of [t^{h}at^h] (Katvalyan 2021: 10).

variance (squared distances between data points and their cluster centers) within each cluster (see Forsyth 2018: 287). Without z-scoring, differences in the scales of the two axes could have caused one or the other axis to have a disproportionate influence on the distances between data points and their cluster centers, resulting in clustering that does not match visually apparent clusters.

After the identification of clusters for each speaker using k-means analysis implemented in R, Euclidean distance between the cluster centers was calculated for each speaker using the original normalized (but not z-scored) data.[14] As in the case of the other measurements discussed above, the Euclidean distances resulting from this clustering analysis were entered as the dependent variable in a fixed-effects linear regression model in R with GENDER, BIRTH YEAR, and EDUCATION LEVEL as predictors. The statistical significance of the predictors in this linear model provides evidence for which demographic characteristics are more likely to be associated with separate pronunciations of [æ] and [a], which in turn implies greater adherence to the dialect vowel system.

2.3.3 Results

2.3.3.1 Maintenance of vowel distinctions

GENDER was a significant predictor of both Pillai scores and ED-median for three of the four vowel comparisons. Model results in relation to GENDER are presented in **Table 2.1**. As BIRTH YEAR and EDUCATION LEVEL were never significant, the model results in relation to these predictors are not presented here (see Portugal [2025] for these results). **Figure 2.5** presents individual participants' median normalized formant measurements for each vowel category, grouped by gender.[15] This figure and all other data visualizations in this chapter were created in R with *ggplot2*.

Table 2.1: Results of linear models in relation to GENDER. P-values of less than 0.05 are italicized. All values are rounded to three decimal places.

[14] An attempt was also made to repeat this analysis using data that had been converted to Bark before being normalized. However, even with z-scoring, using the Bark-converted data set resulted in difficulties for the k-means analysis, such that the clusters identified by k-means did not correspond with clear visual clusters.

[15] This plot excludes some words that were outliers for their categories, which were removed after normalization as described in Portugal (2025), as well as two words intended to elicit the [æ] category for which this categorization was later deemed dubious (see footnote 13). Though omitted from this plot, these data were included in the normalization, as Nearey normalization does not depend on the category identity of the tokens.

Comparison	Pillai scores	ED-median
[a]/[ɒ]A	*p = 0.047* adjusted R-squared = 0.036	*p = 0.041* adjusted R-squared = 0.091
[ɔ]/[ɒ]O	*p < 0.015* adjusted R-squared = 0.103	*p < 0.013* adjusted R-squared = 0.102
[ɒ]A/[ɒ]O	*p < 0.001* adjusted R-squared = 0.328	*p < 0.001* adjusted R-squared = 0.518
[ie]/[ɛ]	p = 0.207 adjusted R-squared = -0.021	p = 0.400 adjusted R-squared = -0.111

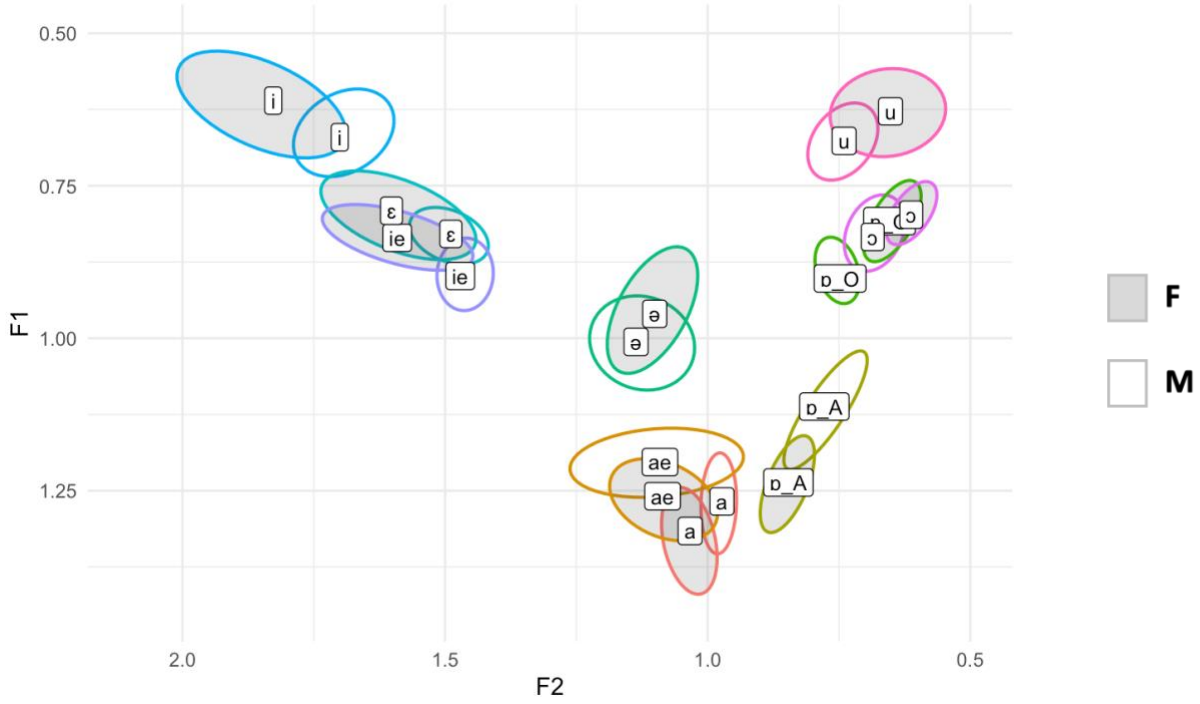


Figure 2.5: Individual median formant values grouped by GENDER in normalized units (without Bark conversion). Confidence ellipses created in *ggplot2* with the *stat_ellipse()* function represent approximately one standard deviation.

In all models where GENDER was significant, the directionality of the differences between gender groups indicated that men were using vowels more consistent with Katvalyan's (2016) descriptions than women. Figures visualizing these differences are available in Portugal (2025), and select figures are also presented in Chapter 3.

Despite the consistent nature of these gender patterns, the reliance of the analysis solely on data from the picture and word tasks limits the interpretation of the results and raises questions about potential stylistic variation that may not have been captured. In order to address this limitation, a small preliminary analysis of the [a]/[ɒ]$_A$ vowel pair in conversational speech from the 2022 interviews was undertaken (see Portugal [2025] for more information about the methods and results of this analysis). This analysis uncovered age-based variation in which older participants used more dialect-like variants, suggesting that older participants may have been more likely to shift toward SEA in the picture and word tasks (thus obscuring any pattern of age-based variation). Nevertheless, the analysis also found a gender pattern similar to the ones in the picture/word data, and did not find any clear education pattern. This suggests that the findings related to these two demographic variables are stable and robust enough to be detected across multiple speech contexts, despite the possible presence of some stylistic variation.

In general, the results suggested that some speakers are maintaining some of the dialect vowels described by Katvalyan (2016). However, many speakers also used a vowel system similar to that of SEA and the colloquial variety of Yerevan. It should be noted that the results of this study combined with Katvalyan's descriptions cannot tell us clearly whether the identified variation represents a change in progress (i.e., ongoing dialect leveling) or stable variation. While the general lack of age-based variation would suggest that it represents stable variation, the results of the preliminary analysis of spontaneous speech indicated that there may exist age-based variation that was not captured in the main analysis. Additionally, it is possible that the variation observed in the 2022 results is stable, but nevertheless represents a change from the earlier dialect vowel system described by Katvalyan (2016), which may not have been characteristic of all or most residents of Gavaṛ for many decades already. Speakers' partial adherence to the dialect vowel system conflicts with Katvalyan's statement that there are not phonetic and grammatical differences between the Bayazet dialect of the late 19th and early 20th centuries and the dialect of today (2016: 19). However, this discrepancy can likely be explained by the differing methods of dialectological and variationist studies. In particular, it is possible that Katvalyan's work seeks out the most conservative speakers – generally "nonmobile, older, rural males" (Chambers & Trudgill 1998: 29). Accordingly, perhaps most or all participants of the 2022 study would not be considered dialect speakers within Katvalyan's framework. As Maguire (2007: 89-90, 165-166) points out, traditional dialectologists go to great lengths in order

to find speakers who use conservative dialect forms, and their data is not necessarily intended to represent the speech community as a whole. It is thus necessary to be careful when interpreting such data (Maguire 2007: 90) and to keep in mind that differences in different data sets may reflect differing methodologies, rather than linguistic change (Maguire 2007: 322). Accordingly, while the results of the 2022 study conflict with Katvalyan's descriptions, it is unclear whether they represent ongoing sound change or whether they simply capture variation that Katvalyan's works do not have the goal of capturing.

The back rounded vowels, for which gender-based variation was found, were also the vowels most often mentioned in participants' metalinguistic commentary (with [ɒ]$_A$ in particular being referenced in all cases where it was possible to discern which specific back rounded vowel the participant had in mind). The vowel [ɒ]$_A$ is also the most salient aspect of the dialect vowel system according to Katvalyan (2016: 80). In contrast, no demographic variation and very limited metalinguistic commentary was found in relation to [ie]/[ɛ]. The results thus suggest a connection between salience of vowels and their maintenance.

Pillai scores and ED-median values for [ie]/[ɛ] were lower overall than for the other comparisons, suggesting that this distinction is being lost. Examining formant trajectories may have provided more insights into differences between the two categories, given that [ie] is supposed to be diphthongal. However, none of the analyzed (picture and word list) tokens seemed diphthongal based on auditory impressions. Formant trajectories of these tokens were also evaluated visually by plotting measurements taken from the 20-40%, 40-60%, and 60-80% duration intervals individually for each token. Based on this visual analysis, the tokens where [ie] was expected did not generally have an obviously more diphthongal trajectory than those where [ɛ] was expected, and patterns in the trajectory across the three time points were often explainable by potential coarticulatory influence from neighboring consonants in the case of both vowel classes.

Some instances of an audibly diphthongized [ie]-like vowel were observed sporadically during the conversation portions of the interviews of four male participants. One of these participants also diphthongized the vowel in one word in the picture task (which was not included in the main analysis because it was intended to elicit consonant and lexical variation, rather than vowel variation – see Portugal [2025] for a full overview of which tokens were included). However, some of the given words were those in which [ɛ], rather than [ie], is

expected (e.g., *kɛs* [‘half’], *χɛɾ* [‘father;’ elicited in the picture task; compare with SEA *hajɾ*], *pɛtakan*[16] [‘state’ [adj.]]). The question of whether speakers are actually using the [ie] vowel (in words where it is expected, or otherwise) is further complicated by the presence of palatalization of velar consonants before front vowels ([ie], [ɛ], [i], [æ]) in the dialect, which Katvalyan describes in various sources (e.g., Katvalyan 2016: 100; 2025: 89-92). Though Katvalyan (2025: 91, 297) does not list Gavaṛ among those settlements where this feature is found (instead listing only a few of the villages surrounding Gavaṛ), the picture and word-list data of the 2022 study included pronunciations with palatalized-sounding velars before [ie]/[ɛ] (that is, in words where either of these two vowels was expected), and especially before [æ]. Thus, in some cases where the vowel seems to be diphthongal (e.g., the instance of *kɛs* mentioned above, which came from fast spontaneous speech), it is difficult to differentiate the vowel quality from potential coarticulatory influence of a neighboring palatalized velar consonant. Katvalyan does not discuss palatalization of [χ] (which is generally described as velar [x] in Armenian-language sources [see Dum-Tragut 2009: 17], but as uvular in many English-language sources [see Seyfarth et al. 2023]). However, one wonders if the impression of diphthongization in the word *χɛɾ* could also be attributable to palatalization of the consonant. In any case, the overall sporadic use of the diphthongized vowel, and its appearance in words (such as *pɛtakan*) where it is explicitly not expected according to Katvalyan (2016) suggest that the distinction and former distribution of [ie]/[ɛ] is disappearing in Gavaṛ, though the [ie] variant may perhaps still index association with the dialect.

2.3.3.2 Variation in the use of [æ]

Figures 2.6-2.7 show the individual Euclidean distances between the two clusters identified by k-means analysis of the tokens where [æ] and [a] were predicted. **Figure 2.6** presents these individual distances color-coded by GENDER, while **Figure 2.7** presents the data coded by EDUCATION LEVEL. In the linear regression model that included the entire data set presented in these figures, no demographic predictors were significant. However, given that the participant with the largest Euclidean distance (Speaker 30) appears to be an outlier, having a value that increases sharply in comparison with the next highest value, a model that excluded this

[16] Note that the vowel [ie] is expected in the dialect pronunciation of the word *pɛt* (*piet*), which was included in the word list, and from which this adjective is derived. Katvalyan (2016: 93) lists the word *pɛtakan* as an example of a derived word in which the vowel [ie] is not maintained.

outlier was also run. The results of this model are presented in **Table 2.2**. In this model, both GENDER and EDUCATION LEVEL were found to be significant predictors, such that men had larger Euclidean distances than women, and participants with secondary education had larger Euclidean distances than the other two education groups (the p-value for the comparison of the secondary and vocational education groups was obtained using the *contrast()* function from the *contrast* package [O'Callaghan et al. 2022]). These results indicate that men and people with secondary education are maintaining a more dialect-like pronunciation of [æ] (as separate from [a]).

The discrepancy between this result and the results related to the other vowel pairs analyzed above can perhaps be explained by the fact that [æ] and [a] appear to be undergoing a merger by transfer, at least among some participants (see Section 2.3.2.3 above). As noted above, mergers by transfer are common in situations where speakers of a dialect are shifting toward a standard variety and are thought to occur above the level of consciousness. The vowel [æ] was mentioned in some participants' metalinguistic commentary/imitations, suggesting that at least some speakers are conscious of its association with the dialect. The fact that [æ] and [a] are in phonemic contrast (unlike, for example, [a]/[ɒ]$_{A}$) may also mean that speakers have more conscious control over this feature than over those that are either not contrastive or not salient as dialect features. If speakers are thus more conscious of variation and change in this feature and at the same time are more able to consciously control it in comparison with other dialect vowels, it is unsurprising that education-based variation was found in this analysis. It is possible that the secondary education group maintained [æ] to a greater degree due to less exposure to SEA / less orientation toward supralocal varieties. Though the education pattern in this data must be viewed cautiously due to the small per-group sample sizes, a comparison of the 2022 and 2025 production data (discussed in Section 2.6.3 below) suggests that [æ] exhibits style-shifting based on speech context, while evidence for style-shifting is less clear for the other vowel pairs. In light of this, the finding of variation related to EDUCATION LEVEL aligns with the results of many well-known sociolinguistic studies (e.g., Labov 1972), in which individual speakers exhibit stylistic variation in relation to those features that show stratification by socioeconomic status at the population level.

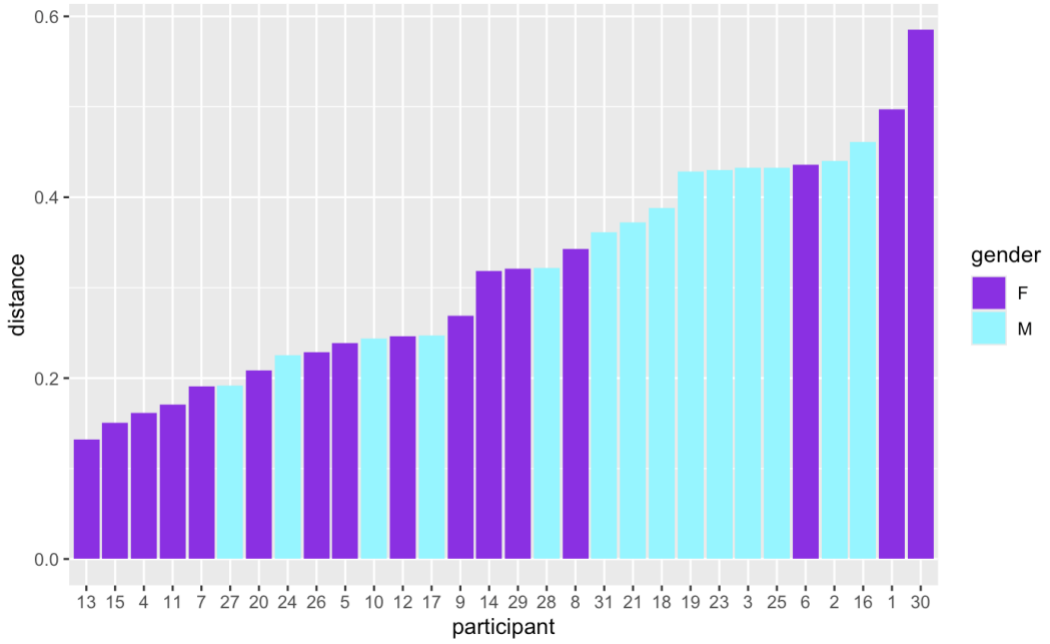


Figure 2.6: Individual Euclidean distance measurements between the centers of clusters identified by k-means analysis of the tokens where [æ] and [a] were predicted. Measurements are presented in normalized units and color-coded by GENDER.

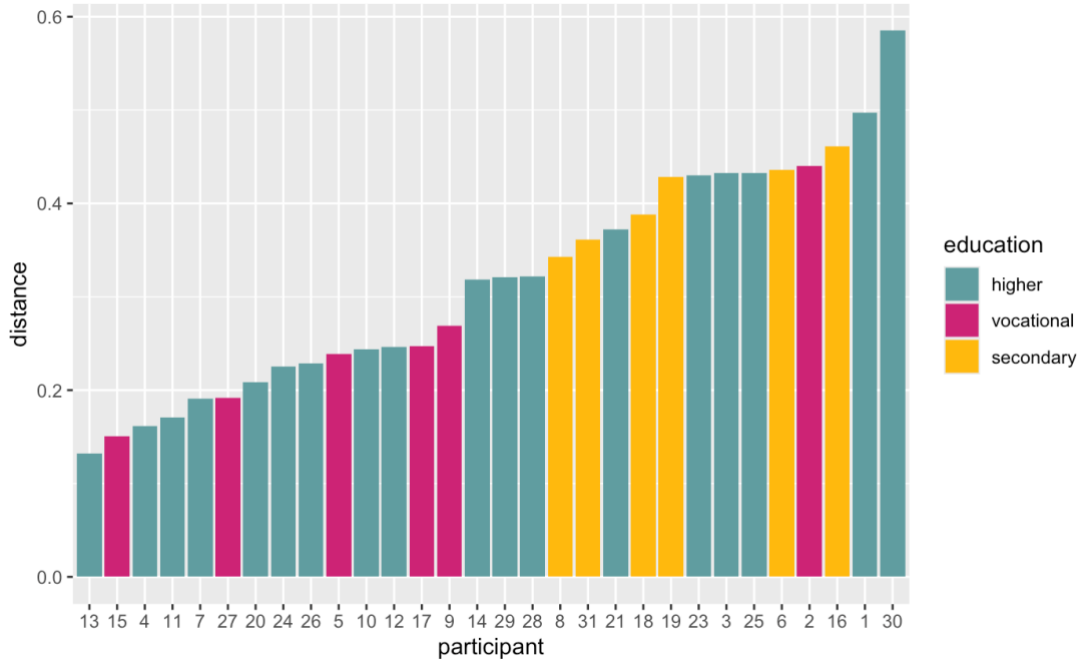


Figure 2.7: Individual Euclidean distance measurements between the centers of clusters identified by k-means analysis of the tokens where [æ] and [a] were predicted. Measurements are presented in normalized units and color-coded by EDUCATION LEVEL.

Table 2.2: Results of the linear regression model predicting Euclidean distance between cluster centers in the analysis of [æ] and [a]. The outlier (Speaker 30) has been excluded from this model. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. The adjusted R-squared of the model was 0.268.

Predictor	**p-value**
GENDER	*0.046*
BIRTH YEAR	0.586
EDUCATION LEVEL – vocational vs. higher	0.590
EDUCATION LEVEL – secondary vs. higher	*0.046*
EDUCATION LEVEL – vocational vs. secondary	*0.031*

2.4 Methods of 2025 production study

2.4.1 Participants

The 51 participants who completed the 2025 study included 14 participants who also participated in the 2022 study (8 women and 6 men). Like the participants of the 2022 study, all participants of the 2025 study were born in Gavaṛ. Six participants (three women and three men) were from the Hats^{h}aṛat neighborhood of Gavaṛ, which was previously considered a separate village, but officially became part of the city in 1962 (Katvalyan 2018: 186). In listing sub-dialects of the Gavaṛ/Bayazet dialect, Katvalyan (2018: 58) lists Hats^{h}aṛat and the city of Gavaṛ in general as belonging to the same sub-group.

Responses to the demographic questionnaire indicated that most participants spent their entire childhood and adolescence in Gavaṛ, with the exception of the participant from 2022 who spent part of his childhood in Kazakhstan (see Portugal 2025), and another participant who lived in Ukraine from ages 3-16. Both of these participants' parents were from Gavaṛ, as were the parents of most of the participants in general. Two participants listed both parents as being from the villages surrounding Gavaṛ, while four participants had one parent born in other cities of Armenia (Yerevan [father], Spitak [mother], Vedi [father], Martuni [mother]). Information was also collected about whether participants had lived outside of Gavaṛ as adults. Seven participants reported having spent time in Russia as adults, while three participants noted having lived in other parts of Armenia or Nagorno-Karabakh (Artsakh). It can be presumed that having spent time in other parts of Armenia / Nagorno-Karabakh was underreported among men, as most of

the men who participated seem to have not listed their military service in the section of the questionnaire that asked about whether they had ever lived outside of their birthplace. One participant listed that they had spent 1.5 years in Russia, but did not list how old they were at that time, so it is unknown if that participant lived in Russia as a child or as an adult. Information was also collected about participants' spouses, in case having a spouse from other parts of Armenia might be found to affect their language use. However, very few participants had spouses from elsewhere, with four participants having spouses from the villages surrounding Gavaṛ and two participants having spouses from elsewhere in Armenia. The four participants with spouses from villages were men, while the two participants with spouses from elsewhere in Armenia were women.

The participant pool was approximately balanced in terms of gender and education level (see **Table 2.3**); however, as participants with vocational education in the 2025 sample were all men, the secondary and vocational groups were pooled into a single "non-higher" group in all analyses. As **Table 2.3** shows, the two educational groups were somewhat unbalanced within each gender group; however, this imbalance may be reflective of actual education trends among the Armenian population. In 2023, the ratio of women to men enrolled in tertiary (higher) education in Armenia was 1.27, indicating a higher proportion of women among tertiary-education students (TheGlobalEconomy.com 2023). Participants ranged in age from 18 to 50 (see **Figure 2.8**) and had self-reported normal hearing. An upper age limit of 50 years old was chosen in order to recruit participants who would be less likely to have issues with vision and hearing, and who would be more likely to feel comfortable using unfamiliar computer software to complete the experiment (this age threshold was determined in collaboration with the local research assistant, based on her impressions about which types of people might experience difficulty with the experimental procedure).

Table 2.3: Summary of participant information with respect to GENDER and EDUCATION LEVEL.

	Higher	Non-higher	Total
Men	11	14	25
Women	18	8	26
Total	29	22	

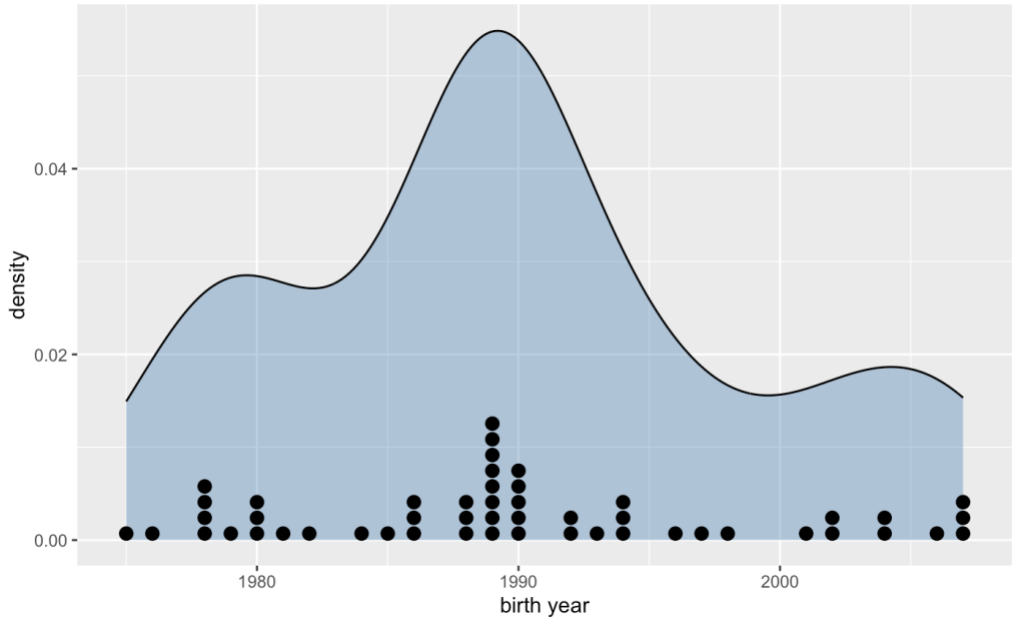


Figure 2.8: Density of participants by birth year. Each dot represents one participant.

2.4.2 Materials and data collection

The word list included words containing both the target vowels ([a]/[ɒ]$_{A}$ and [ɔ]/[ɒ]$_{O}$, which were the same vowel pairs investigated in the perception tasks) and the remaining vowels of Armenian, which acted as fillers and were also necessary for the normalization of the formant data (which is most effective when the entire vowel space is included). Like in the 2022 study, the words were not embedded into a carrier sentence, as full written sentences would have been more likely than individual words to prompt speakers to shift toward a more standard style.

All target vowels, as well as all additional vowels whose formants were measured to fill out the vowel space, were under either primary or secondary stress. It was not considered necessary for all vowels to be under primary stress, as Haghverdi (2016) has found that Armenian stressed and unstressed syllables have no significant length difference. In Gavaṛ's dialect and in SEA, the last full vowel (i.e., not schwa) of a word is stressed, while the first syllable has secondary stress (Abeghyan 1971: 19-20; Allen 1950: 182-183; Dum-Tragut 2009: 47; Katvalyan 2016: 104; Khach^{h}atryan 1988: 80; Vaux 1998: 134).

The target vowels were represented in the word list primarily by near-minimal pairs. Although the [ɔ]/[ɒ]$_{O}$ distinction is phonemic according to Katvalyan (2016), there were no minimal pairs suitable for use in this study, as the pairs listed by Katvalyan include words that

are not equal in their stylistic associations (e.g., one word is shared with SEA, while the other is strictly a dialectal word, which would be unusual to present in writing).

Words containing the target vowels included both the words used as stimuli in the perception tasks and words forming near-minimal pairs with those words, as well as some additional near-minimal pairs not used as perception stimuli and some additional words that did not form near-minimal pairs with other words in the list. Given the nature of the environments in which the various target vowels are found, the list was made with a very broad definition of "near-minimal pair" in mind, such that the place of articulation preceding and following the vowel was held constant when possible, while nasality, voicing, and manner of articulation were allowed to vary. For example, the words [mɒs]/[mas] and [bad] (see the selection of words in **Table 2.4**) were considered a near-minimal pair because the vowel is followed by the same place of articulation (dental) in both words, while the preceding consonant is bilabial in both words. As the features of the preceding consonants (i.e., being nasal vs. being voiced and non-nasal) are what defines the phonetic environment for the target vowels (allophones), they did not disqualify these words from being considered a near-minimal pair. Likewise, the words [χɒz]/[χɔz] and [ɔt^{h}] were considered a near-minimal pair because the vowel is followed by the same place of articulation and preceded by an environment (in this case, having vs. not having any preceding consonant) that determines which vowel will appear (as the lack of any preceding consonant means that [ɔ] will appear – see Section 1.3.2.2 above). While the word list did contain some actual minimal pairs (e.g., the words [tɒs]/[tas] and [das]), in which a single contrastive segment differentiates the meaning of the two words, none of these were cases in which the vowel itself was the contrastive segment.

As **Table 2.4** shows, the pairs containing [a]/[ɒ]$_{A}$ more frequently held the preceding and following place of articulation constant across the two words in comparison with the words containing [ɔ]/[ɒ]$_{O}$. Indeed, the only possible monosyllabic near-minimal pairs that could be identified to elicit [ɔ]/[ɒ]$_{O}$ do not hold the preceding and following places of articulation constant because the alternation of these vowels in monosyllabic words usually depends on the presence vs. absence of a preceding consonant, or on the following consonant being [v] vs. any other consonant. It is possible to attempt to elicit the [ɔ]/[ɒ]$_{O}$ difference with words in which it arises due to movement of the stress in word derivation (e.g., *g^{h}ɒʁ* 'thief' vs. *g^{h}ɔʁut^{h}yun* 'theft'). However, perception stimuli could not be successfully created using the talker's pronunciations

of these words, which suggested that the difference between the vowels in such pairs may generally not be as large as in some other words that could be used to elicit them. Thus, the list ultimately included many pairs in which the preceding and following places of articulation were not held constant.

Most words containing the target vowels had four repetitions throughout the list. All other words (including some words with the target vowels, as well as words making up the rest of the vowel space) occurred only once or twice in the list. The full word list (in the order of presentation) is provided in Appendix A. The list contained 186 words in total (including all repetitions).

Table 2.4: Selected words from the word list. Those words used as perception stimuli in either perception task, along with their near-minimal pairs that were used in the creation of the perception stimuli, are bolded. Note that the 2025 word list also contained the word [baɾtʰ] ('complex') as a more accurate pair for [mɒɾtʰ]/[maɾtʰ], but this word could not be used in the creation of perception stimuli, as it was not included in the 2022 word list.

comparison	Predicted dialect pronunciations of pair	SEA pronunciations of pair
[a]/[ɒ]A	**[pɒt] ('wall') and [bad] ('duck')**	**[pat] ('wall') and [bad] ('duck')**
	[mɒs] ('part') and [bad] ('duck')	**[mas] ('part') and [bad] ('duck')**
	[mɒɾtʰ] ('person') and [bad] ('duck')	**[maɾtʰ] ('person') and [bad] ('duck')**
	[tɒs] ('ten') and [dʰas] ('lesson')	[tas] ('ten') and [das] ('lesson')
	[tɒr] ('letter') and [dʰaɾ] ('century')	[tar] ('letter') and [daɾ] ('century')
[ɔ]/[ɒ]O	**[t͡sɒt͡sʰ] ('bosom, embrace') and [ɔtʰ] ('air')**	**[t͡sɔt͡sʰ] ('bosom, embrace') and [ɔtʰ] ('air')**
	[pʰɒs] ('hole') and [ɔtʰ] ('air')	**[pʰɔs] ('hole') and [ɔtʰ] ('air')**
	[χɒz] ('pig') and [ɔtʰ] ('air')	**[χɔz] ('pig') and [ɔtʰ] ('air')**
	[bʰɒt͡sʰ] ('flame') and [ɔt͡sʰ] ('snake')	[bɔt͡sʰ] ('flame') and [ɔt͡sʰ] ('snake')
	[gʰɒʁ] ('thief') and [gʰɔʁutʰyun] ('theft')	[gɔʁ] ('thief') and [gɔʁutʰyun] ('theft')

Data collection was carried out in June of 2025 with the help of the same local research assistant who conducted the sociolinguistic interviews in 2022. All participants interacted with both me and the research assistant. The study was carried out in a quiet room at the Geghark^{h}unik^{h} Regional Library in Gavaṛ (see Section 3.5 for a full overview of the data collection procedures [including both the production and perception tasks]). There were two participants (Participants 39 and 40) who completed the study in a quiet room at a private home, due to difficulties in accommodating their work schedules. As in the 2022 study, participants were instructed to read the words as they would usually say them in everyday speech. This unconventional elicitation method was chosen due to relatively high metalinguistic awareness about the existence of different varieties of Armenian. During the recording of the word list, only the research assistant remained in the room with the participant, both so that background noise would be minimized and so that participants would be less likely to alter their speech in the presence of a non-local person. The word list was displayed using Microsoft PowerPoint slides, with words being presented one at a time on separate slides. The research assistant advanced the slides while the participant read the word list. Recordings were made with a Zoom H1 XLR audio recorder (a 32-bit float recorder). The recorder was set to a sampling rate of 48,000 Hz. Participants were recorded with an Audio-Technica AT803 omnidirectional condenser lavalier microphone. The recorder battery started to run low during one participant's recording (Participant 40), which negatively affected the quality of this recording. As such, this participant (who had originally completed the recording and experiment in a private home) was recorded reading the word list for a second time the next day at the Geghark^{h}unik^{h} Regional Library.

2.4.3 Analysis

2.4.3.1 Formant measurements and analysis of maintenance of vowel distinctions

Segmentation of vowel tokens was performed manually in Praat (Boersma & Weenink 2021). Thomas (2011: 139) discusses that one can make various decisions in determining where to mark the boundaries of a vowel, but stresses the importance of being consistent across all tokens. In the present study, a general attempt was made to follow some of Thomas's recommendations – specifically, an attempt was made to mark the start of the vowel at the "onset of glottal pulsing for F2" (Thomas 2011: 140) and to mark the end of the vowel at the point where the glottal pulsing of F2 disappeared or became indistinct (which happens with varying

abruptness depending on the following consonant). Despite this effort to maintain consistency, it is acknowledged that segmentation ultimately involves subjective decisions due to vast variation both within and across speakers in terms of how gradient or abrupt the onset and offset of glottal pulsing appears on the spectrogram (which, in the present data set, appears to result not only from identifiable factors such as the sporadic use of breathy voicing among some speakers, but also from individual idiosyncrasies). The difficulties encountered during segmentation align with the discussion presented by Maguire (2007: 210-214), who stresses that vowel segmentation necessarily involves subjective decisions (especially with sub-optimal recording quality) and argues (2007: 225) that issues of subjectivity and strict replicability are "inherent in acoustic analysis, and to pretend otherwise is to assume a rigour which may not be justified."

The first and second formants were measured from each participant's recording using Fast Track (Barreda 2021). Fast Track took measurements every 2 milliseconds throughout the 40-60% duration interval of each vowel token. For each token, the median formant value among these measurements was recorded for use in further analyses. Measurements from the 20-40% and 60-80% duration intervals were also collected, though these are not included in the main analysis (see Section 2.4.3.2 below for discussion of their use in investigation of diphthongization). The formant measurements for each token were visually checked using spectrograms automatically generated by Fast Track. Visually checking each token is preferable to simply checking/removing outliers because it allows for the correction of small inaccuracies (Kendall & Fridland 2021: 48), which are more difficult to detect than outliers (Labov 2001: 154). Those measurements that appeared inaccurate were corrected either by selecting other measurements generated by Fast Track (which generates 20 different analyses for each token using a range of settings for the maximum frequency below which it will identify formants), or by replicating Fast Track's procedure in Praat. Appropriate settings in Praat (maximum frequency and number of formants) were determined visually for each token (i.e., by visually determining which combination of settings best corresponded to where the formants appeared to be on the spectrogram). Praat's formant measurements were then recorded at intervals of 2 milliseconds throughout the 40-60% duration interval of the vowel token, and the median of these values was then calculated.

For the participant for whom two recordings were made due to the recorder battery running out during the initial recording (Participant 40), measurements were taken from the

second recording when possible, but were taken from the first recording when possible in cases where accurate formant tracking could not be obtained from the second recording. This participant's voice was quiet and creaky, meaning that formants in the second recording were sometimes less distinct than in the first recording despite the battery running low during the first recording.

Some tokens with indistinct formants were excluded prior to measurement (e.g., when it was apparent during segmenting that the formants would be too indistinct to measure). Each participant's raw formant data was visualized via individual vowel plots, so that the formant measurements for tokens appearing to be outliers (for their categories or in general) could be checked. Tokens for which accurate formant tracking could not be obtained with either Fast Track or Praat were excluded. Two tokens were excluded from the data set before normalization for reasons other than bad formant tracking – one of these appeared to be an extreme outlier for the given participant's vowel space and thus could have skewed the normalization, while the other was an extreme outlier for its vowel category in the given participant's speech and was deemed by the research assistant to be an artificial performance of dialect pronunciation.[17] A summary of sample sizes is provided in **Table 2.5**, which shows the sample sizes of the speakers who had the maximum, minimum, and median amount of tokens for each vowel category and in total. The table reflects sample sizes after the exclusion of outliers. The maximum number of total tokens exceeds the total number of words in the list (186) because some words (e.g., two-syllable words) contained two vowels that were measured (there were also occasional instances in which participants repeated a word, thus providing additional tokens that could be measured).

Table 2.5: Median, maximum, and minimum sample sizes among speakers by vowel category and in total. Categories are grouped into those that were and were not included in the analysis.

Vowel	Median	Maximum	Minimum	Included in analysis?
[a]	39	42	37	Yes
[ɒ]A	40	42	33	Yes
[ɒ]O	38	40	37	Yes
[ɔ]	25	27	24	Yes

[17] This participant had also participated in the 2022 study, and some seemingly artificial performances of dialect pronunciations were also noted in that data based on the research assistant's impressions about the participant's usual everyday speech, though the instance in the 2025 data was more extreme than anything noted in the 2022 data.

[æ]	7	8	6	No
[ɛ]	9	10	8	No
[ie]	12	14	12	No
[i]	10	11	8	No
[u]	11	12	9	No
total	191	196	183	

As normalized data would be required for some of the further analyses, the 40-60% interval data was normalized using the Nearey1 method from the *vowels* package (Kendall & Thomas 2018). This method was chosen because it does not require F3 and has been found to perform well at preserving social variation (Thomas 2011: 167-168). Among Nearey's two normalization methods, the formant-intrinsic Nearey1 method has been found to perform better at preserving social variation while reducing physiological variation (Adank et al. 2004). Like the 2022 data, the data was normalized both with and without first converting Hz to Bark. As the focus of the perception experiment is the back rounded vowels of Gavaṛ's dialect, Pillai scores and ED-median measurements were only calculated for three of the vowel pairs investigated in the 2022 study – [a]/[ɒ]$_A$, [ɔ]/[ɒ]$_O$, and [ɒ]$_A$/[ɒ]$_O$. As noted above, less overlap and greater distance between the two vowels equates to greater adherence to the dialect system for the first of these two pairs, while more overlap and smaller distance between the two vowels equates to greater adherence to the dialect system for the third pair. Pillai scores and ED-median were calculated following the same procedure described in Section 2.3.2.2.1. As mentioned above, ED-median values differed depending on whether the data was converted to Bark prior to normalization. In cases where the results of analyses were the same with or without conversion to Bark, results will be presented without Bark conversion for the sake of consistency with the presentation of the 2022 results in Portugal (2025). However, in cases where the results differ with and without Bark conversion, both sets of results will be presented below.

As in the 2022 analysis, data visualizations and fixed-effects linear regression models were used to investigate which demographic predictors may be associated with adherence to the dialect vowel system. Separate models were run for each vowel pair and for each dependent variable (Pillai scores and ED-median). GENDER was once again implemented as a categorical predictor with two levels, while EDUCATION LEVEL was also represented by a categorical

predictor with two levels (higher and non-higher). BIRTH YEAR was represented by a continuous predictor. These three predictors were included in all models because they represent key macrosocial factors frequently investigated in production data in sociophonetic research (see Section 2.3.2.2.2 above). In order to explore other trends that may have been present in the data, two additional predictors were also tested – PREVIOUS and HATSARAT. These predictors represented whether participants had participated in the 2022 study and whether participants were from the Hats^{h}aṛat neighborhood of Gavaṛ. It was expected that participants who had completed the 2022 study might have productions that were different from those of new participants, either due to general sampling issues or due to potentially feeling more comfortable being recorded. Though Hats^{h}aṛat is officially part of Gavaṛ, it was also expected that participants from this neighborhood might differ from other participants in their productions due to its history of being a separate village. In addition to these two predictors, interactions between different demographic predictors were also tested. However, interactions with the HATSARAT predictor were not tested due to the small sample size of the group from this neighborhood (n=6).

Different models were compared using two statistical metrics – AIC (Akaike Information Criterion) with correction for small sample size and BIC (Bayesian Information Criterion). AIC and BIC help to determine which model offers the best fit for the given data set without being overly complex (and thus potentially overfitted to the particularities of that data set) (see Pardoe 2020: 259-260). Both metrics were calculated in R. AIC with small sample size correction was implemented using the *AICc()* function from the *AICcmodavg* package (Mazerolle 2023). It was determined that small sample size correction was necessary because each model included only 51 data points. In cases where the two metrics disagreed (which often occurs due to BIC preferring a simpler model – see Emmert-Streib et al. [2023: 315]), or where AIC/BIC values for different models were very close (which in general commonly occurs – see Pardoe [2020: 260]), the results of multiple models are presented and/or discussed. Results of the best models are presented along with data visualizations that illustrate trends in relation to various predictors and their interactions.

2.4.3.2 Diphthongization

As the vowels [ɒ]$_{A}$ and [ɒ]$_{O}$ are diphthongal according to previous dialect descriptions, (e.g., Katvalyan 2016), various methods for evaluating potential diphthongization of these

vowels in the 2025 study were considered. Nevertheless, as in the 2022 study, the overwhelming majority of tokens did not seem to be obviously diphthongized in the expected manner based on auditory impressions. Though methods exist for examining dynamic formant trajectories and comparing them across different tokens and vowel classes (e.g., Generalized Additive Mixed Models [see Clopper 2024; Cole & Strycharczuk 2019; Kirkham et al. 2019; Renwick 2017; Renwick & Stanley 2020; Sóskuthy 2017; Warburton 2020] and Discrete Cosine Transforms [see Cox et al. 2024]), it was unclear whether the application of such methods in the present study would provide any additional insights in comparison with the analysis of Pillai scores and ED-median. Since the vast majority of tokens in the present data set did not seem audibly diphthongized in the expected manner, one must wonder whether the subtleties of formant trajectories captured by the given methods would represent meaningful variation between speakers and demographic groups, or if it could be attributed to random noise, coarticulatory effects of surrounding consonants, or other subtle aspects of speech production that are not relevant to the research questions.

Some studies that compare the outcomes of different formant analyses (e.g., Butcher [2022], who uses Pillai scores [calculated for both the vowel nucleus and offglide] and GAMMs) indeed raise questions about the pros and cons of using static vs. dynamic analyses. In comparing which speakers are deemed to have merged vs. unmerged vowels based on nucleus and offglide Pillai scores (using p-values from the MANOVA output) vs. GAMMs (which includes its own significance testing), Butcher finds that "the most commonly occurring misalignment related to participants being categorised as merged according to Pillai measurements, but as distinct by GAMMs criteria" (2022: 184-185). While Butcher argues that such discrepancies justify the use of both methods in examining vowel mergers, it is worth questioning whether methods such as GAMMs may be too sensitive in some cases, potentially capturing variation that is not meaningful (in relation to the goals of the study) in cases where the vowels are not audibly diphthongized. As the presentation and discussion of the results in Portugal (2025), Section 2.5 below, and Chapter 3 suggest, speakers producing vowel differences that seem unlikely to be perceptible are often deemed unmerged by MANOVA p-values, even when adjusted thresholds have been used to account for small sample sizes (following the method of Stanley & Sneller [2023]). Accordingly, it seems that Pillai score p-values may already be too sensitive a measure of vowel differences in some cases, and using metrics that uncover even finer-grained variation

may not be useful or advisable in cases where the subtleties of time-course data are not obviously relevant given the research questions and characteristics of the data set. (Nevertheless, the choice of methods ultimately depends on the goals of the study, and detailed time-course information may be useful in cases where one seeks to examine the subtleties of speech production, even if those subtleties are not perceptible or socially meaningful.)

In light of these issues, a preliminary analysis was undertaken in order to assess whether it would be worth implementing more complex methods of analyzing formant trajectories (such as GAMMs or DCT). As noted above, median formant measurements from the 20-40%, 40-60%, and 60-80% duration intervals were collected for each vowel token. These three measurements were plotted individually for each [ɒ]$_A$ and [ɒ]$_O$ token, and the trajectory across the three measurements was visually examined (see **Figure 2.9** for an example plot). Tokens which appeared to potentially have the expected diphthongization pattern – moving from more [ɔ]-like to more [a]-like pronunciation (i.e., the vowel becomes lower and/or more front across the three time points) – were counted for each participant and compared with that participant's total number of tokens in the given category. An attempt was made to count such tokens in the most inclusive way possible, such that any token that even vaguely resembled the expected pattern would be counted. For comparative purposes, the same procedure was carried out for each participant's [a] and [ɔ] tokens (which are not expected to show the same pattern of diphthongization according to previous dialect descriptions). For each participant, the percent of tokens potentially having the expected diphthongization pattern was then calculated for each of the four categories. **Table 2.6** presents the median, mean, maximum, and minimum percent among participants for each category. As the table shows, it does not seem that tokens potentially having the expected pattern were drastically more prevalent in the [ɒ]$_A$ and [ɒ]$_O$ categories than in the [a] and [ɔ] categories, which suggests that the given trajectory shape may be resulting from factors that would affect all four vowel categories approximately equally (such as coarticulatory effects of neighboring consonants). Given the outcome of this preliminary analysis, it was determined that additional analysis of formant trajectories should not be pursued.

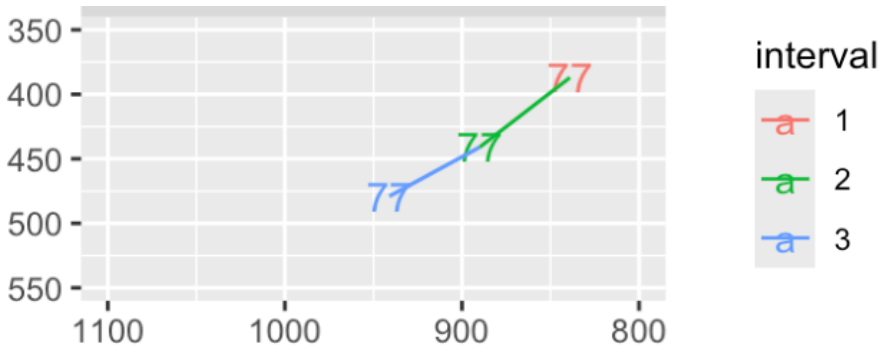


Figure 2.9: A token of [ɒ]O with a formant trajectory that could represent the expected pattern of diphthongization (moving from more [ɔ]-like to more [a]-like). The axes are shown in Hz. The three measurements represent the 20-40%, 40-60%, and 60-80% duration intervals of the vowel token.

Table 2.6: Median, mean, maximum and minimum percentages of tokens in each category that potentially show the expected diphthongization pattern.

vowel	median percent	mean percent	maximum percent	minimum percent
[ɒ]A	18.4%	19.5%	52.5%	0%
[ɒ]O	26.3%	25.3%	50.0%	0%
[a]	21.6%	24.2%	61.5%	2.6%
[ɔ]	16.0%	17.7%	46.2%	0%

2.5 Results

This section presents the 2025 production results separately for each of the three examined vowel pairs ([a]/[ɒ]A, [ɔ]/[ɒ]O, and [ɒ]A/[ɒ]O), reviewing the demographic predictors that were found to be significant in models predicting Pillai scores and ED-median for each pair. The main result of this analysis is that gender-based variation was found for [a]/[ɒ]A and [ɒ]A/[ɒ]O, corresponding to the variation found in 2022. On the other hand, in contrast to the results from 2022, gender-based variation was not found for [ɔ]/[ɒ]O. Differences between previous (i.e., repeat) and new participants were sometimes found and are especially important to consider in the case of [ɔ]/[ɒ]O, where they might explain the differences in gender patterns between the 2022 and 2025 data. **Figures 2.10** and **2.11** below present individual participants' median normalized formant measurements for each vowel category, grouped by GENDER and calculated both without and with Bark conversion.

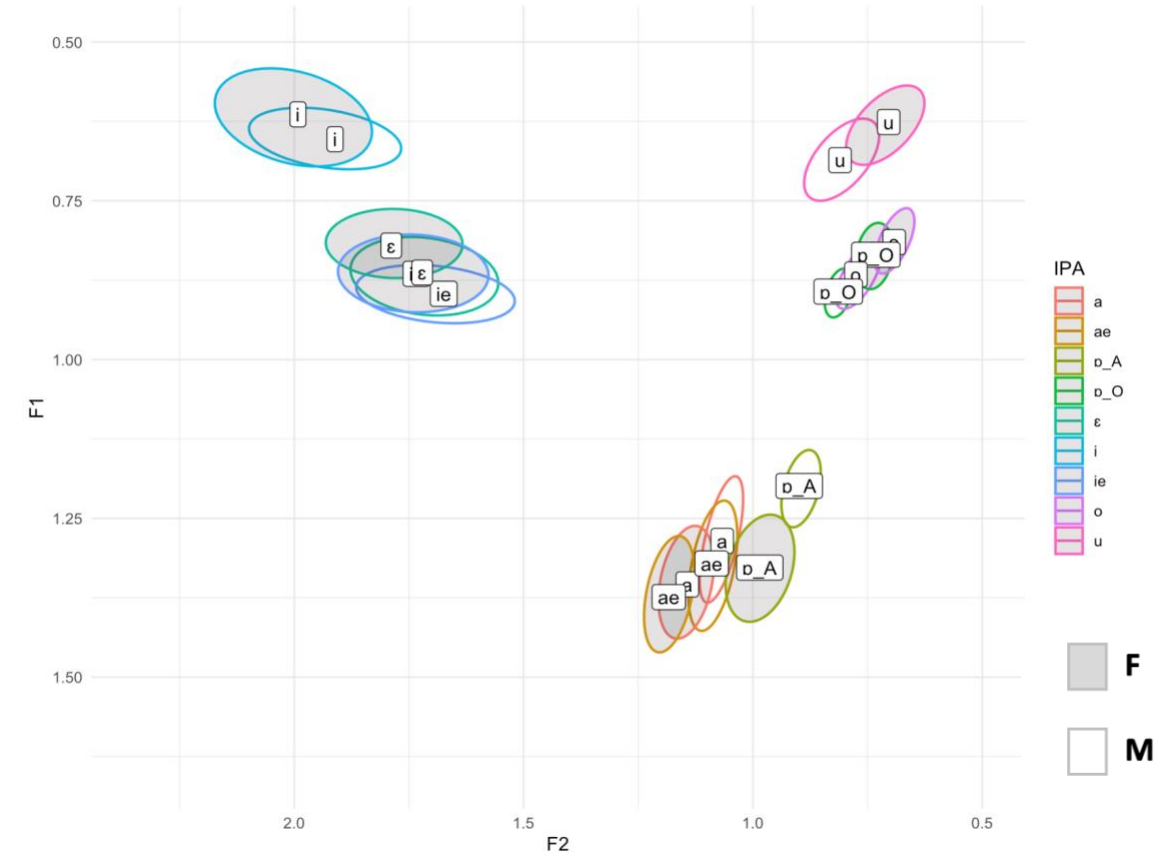


Figure 2.10: Individual median formant values from 2025 grouped by GENDER in normalized units (without Bark conversion). Confidence ellipses created in *ggplot2* with the *stat_ellipse()* function represent approximately one standard deviation.

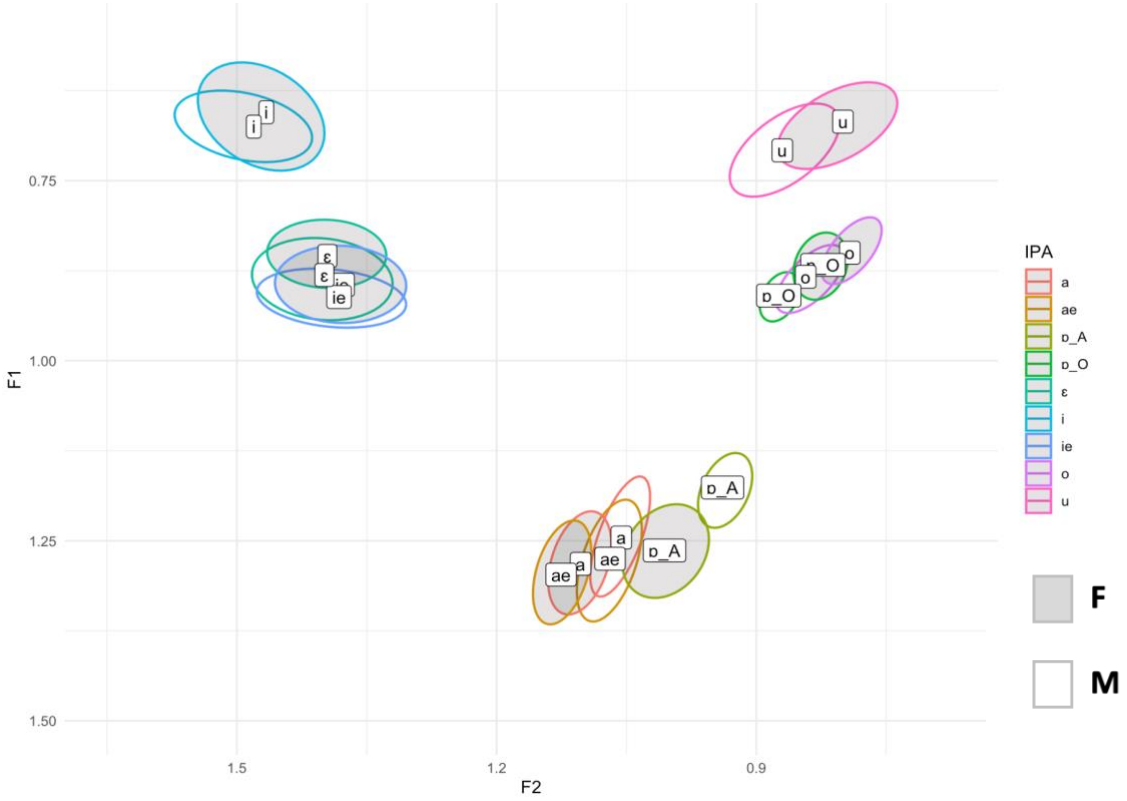


Figure 2.11: Individual median formant values from 2025 grouped by GENDER in normalized units (with Bark conversion). Confidence ellipses created in *ggplot2* with the *stat_ellipse()* function represent approximately one standard deviation.

2.5.1 [a]/[ɒ]$_A$

2.5.1.1 Pillai scores

In models predicting Pillai scores for the [a]/[ɒ]$_A$ comparison, GENDER was found to be a significant predictor, with men having larger Pillai scores (i.e., using more dialect-like vowels) than women. In this regard, the results of the 2025 analysis are similar to those of the 2022 analysis. PREVIOUS came close to significance (p=0.064). AICc preferred the model with PREVIOUS, while BIC (which more harshly penalizes more complex models) preferred the model with only GENDER, BIRTH YEAR, and EDUCATION LEVEL. As a weak pattern in relation to PREVIOUS was nevertheless visible in data visualizations (with previous participants having larger Pillai scores), the results of the model including this predictor are presented. **Figure 2.12** shows Pillai scores for [a]/[ɒ]$_A$ in relation to GENDER, while **Figure 2.13** presents the same data coded in relation to PREVIOUS. Both figures include merger/distinction thresholds calculated with the method of Stanley and Sneller (2023). As the figures show, all participants' vowel categories are deemed distinct by this method, which is similar to the result found in 2022 for this vowel

pair. **Table 2.7** presents the results of the model with GENDER, BIRTH YEAR, EDUCATION LEVEL, and PREVIOUS. The model predicted Pillai scores calculated from non-Bark-converted data, as Bark conversion did not affect Pillai scores (beyond small differences attributable to rounding).

Table 2.7: Results of the linear regression model predicting Pillai scores for [a]/[ɒ]$_A$. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. The adjusted R-squared of the model was 0.102.

Predictor	**p-value**
GENDER	*0.042*
BIRTH YEAR	0.515
EDUCATION LEVEL – non-higher vs. higher	0.644
PREVIOUS	0.064

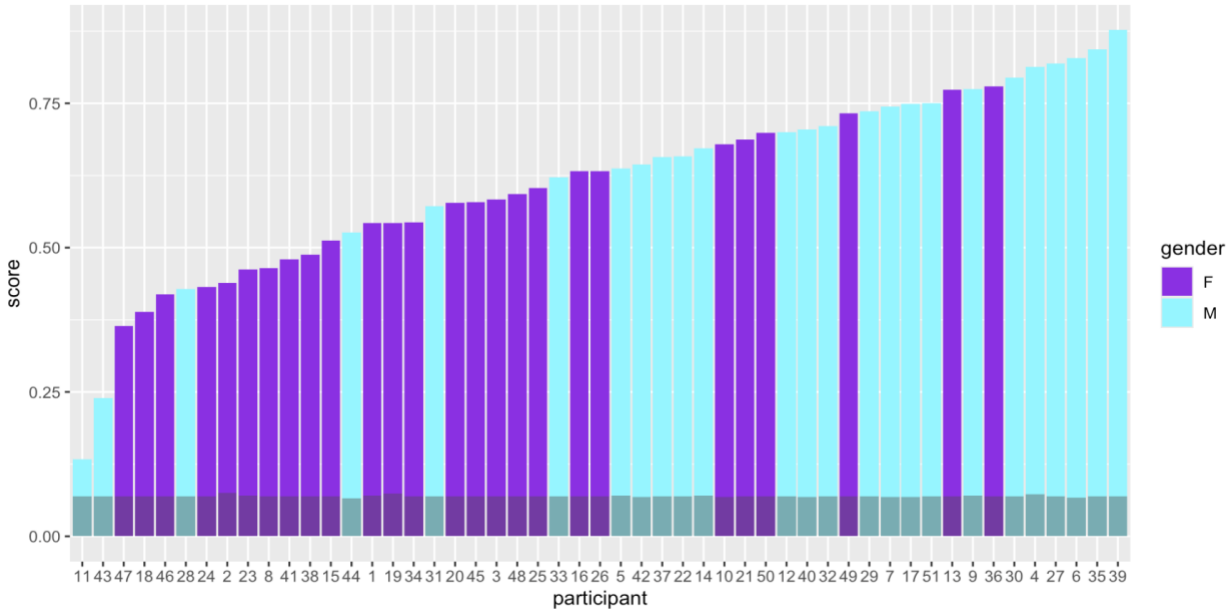


Figure 2.12: Individual Pillai scores by GENDER for **[a]/[ɒ]$_A$** with superimposed merger/distinction thresholds.

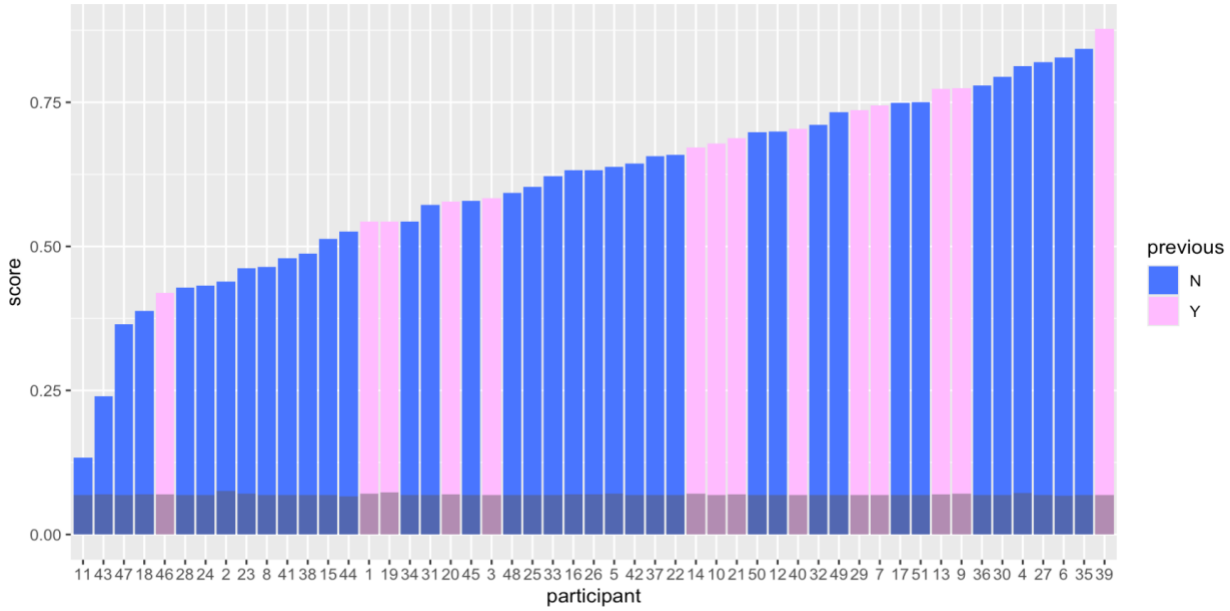


Figure 2.13: Individual Pillai scores for **[a]/[ɒ]$_A$** with superimposed merger/distinction thresholds, coded in accordance with the PREVIOUS predictor.

As **Figures 2.12-2.13** show, there were two participants who were outliers in terms of their [a]/[ɒ]$_A$ Pillai scores, having Pillai scores that were dramatically lower than the next-highest scores (these participants can be seen at the far-left end of the graph). Importantly, both of these participants were men and were not previous participants. **Figures 2.14** and **2.15** present their individual vowel plots. As **Figure 2.14** shows, Participant 11 deviates from the pattern seen in the data of all other 2022 and 2025 participants, in that his [ɒ]$_A$ and [a] clusters do not show any visible difference in the expected direction (in which [ɒ]$_A$ would be further back and also possibly higher than [a]). It is likely that the inclusion of these two outlying participants in the models deflated the statistical significance of GENDER, as the gender pattern seen in **Figure 2.12** is fairly strong if one ignores these two participants. Likewise, the very low Pillai scores of these participants may have inflated the statistical significance of PREVIOUS, which almost reaches significance despite the visible pattern in **Figure 2.13** being quite weak.

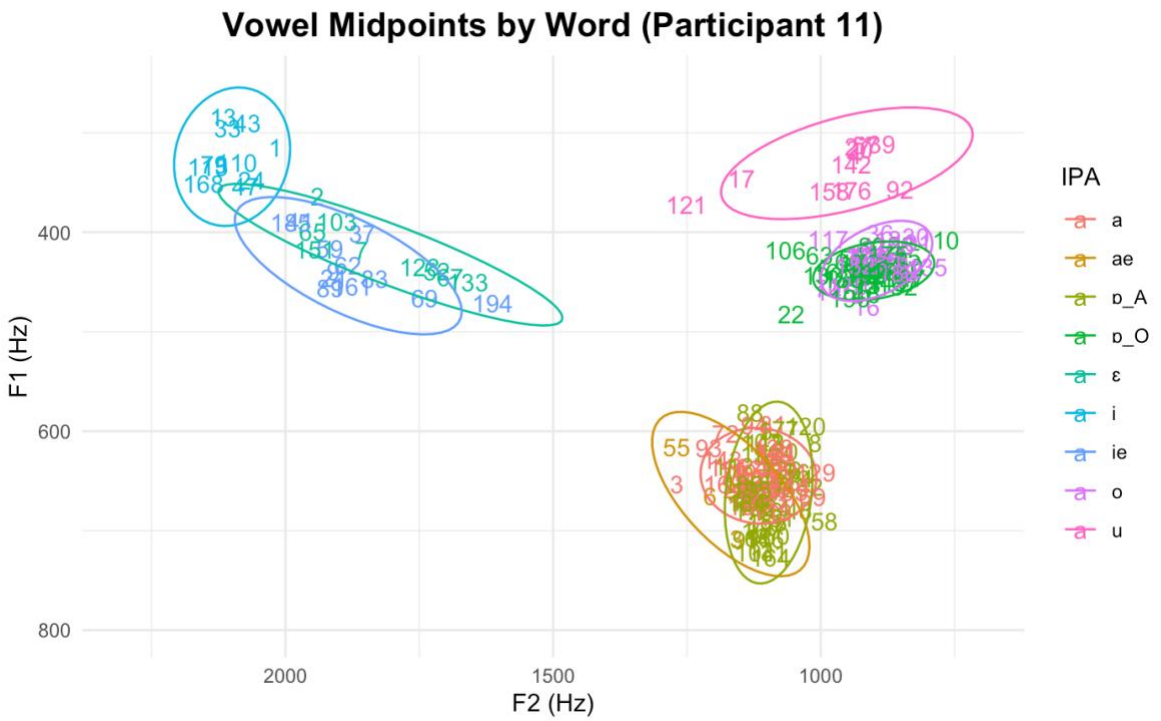


Figure 2.14: Individual vowel plot for Participant 11 (2025). Ellipses represent two standard deviations.

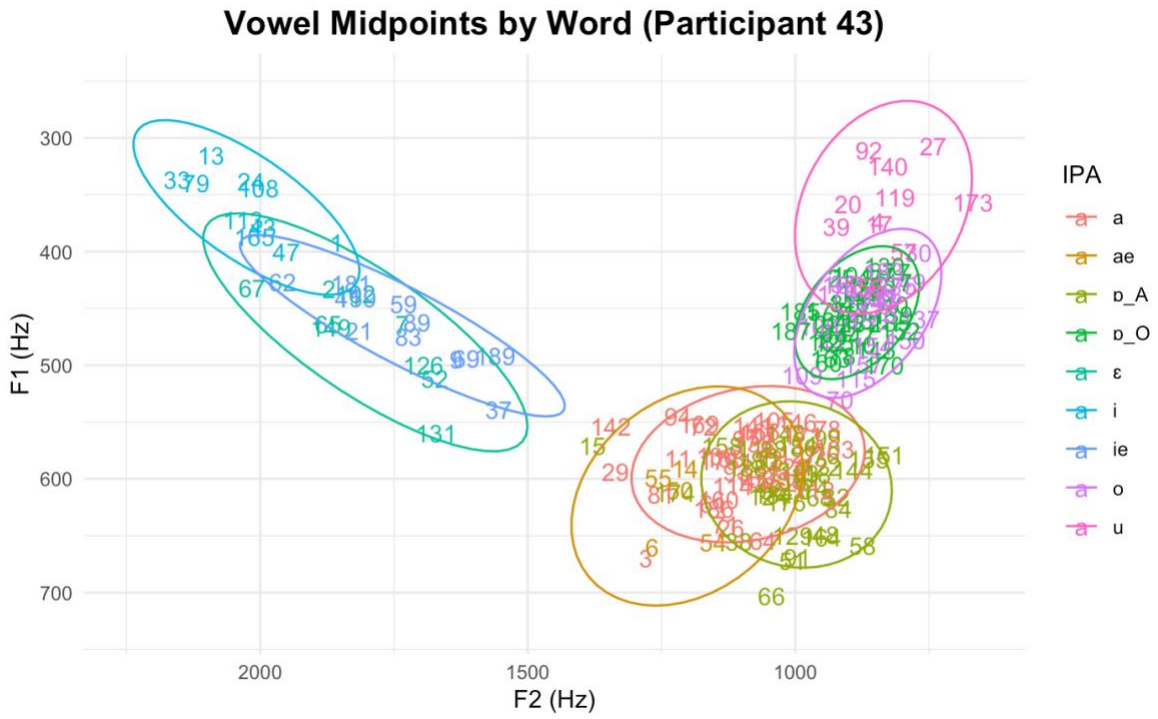


Figure 2.15: Individual vowel plot for Participant 43 (2025). Ellipses represent two standard deviations.

2.5.1.2 ED-median

The results of the models predicting ED-median for [a]/[ɒ]$_A$ differed according to whether or not the data were converted to Bark before being normalized. In models with the non-Bark-converted data, PREVIOUS and the interaction of PREVIOUS with BIRTH YEAR were found to be relevant predictors. AICc preferred the model with the interaction, while BIC preferred the simpler model with only GENDER, BIRTH YEAR, EDUCATION LEVEL, and PREVIOUS. PREVIOUS was significant in this simpler model, while in the model with the interaction, this interaction does not quite reach significance (p=0.082). The results of both models are presented in **Tables 2.8**-**2.9**.

Table 2.8: Results of the linear regression model predicting ED-median for [a]/[ɒ]$_A$ calculated with non-Bark-converted data. The model included PREVIOUS and its interaction with BIRTH YEAR. PREVIOUS was dummy coded (such that the coefficients and p-values for PREVIOUS and BIRTH YEAR do not represent main effects). P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. The adjusted R-squared of the model was 0.089.

Predictor	**p-value**
GENDER	p=0.503
BIRTH YEAR	p=0.461
EDUCATION LEVEL – non-higher vs. higher	p=0.665
PREVIOUS	p=0.081
BIRTH YEAR – PREVIOUS interaction	p=0.082

Table 2.9: Results of the linear regression model predicting ED-median for [a]/[ɒ]$_A$ calculated with non-Bark-converted data. The model included PREVIOUS. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. The adjusted R-squared of the model was 0.046.

Predictor	**p-value**
GENDER	p=0.392
BIRTH YEAR	p=0.997
EDUCATION LEVEL – non-higher vs. higher	p=0.292
PREVIOUS	*p=0.034*

Figure 2.16 presents the ED-median values for this comparison in relation to PREVIOUS (showing that previous participants had more distinct [dialect-like] productions), while **Figure 2.17** illustrates the interaction between PREVIOUS and BIRTH YEAR, according to which the effect of age differs for repeat and new participants. Though GENDER was not significant in these models, **Figure 2.18** presents the data in relation to GENDER, in order to illustrate the lack of pattern.

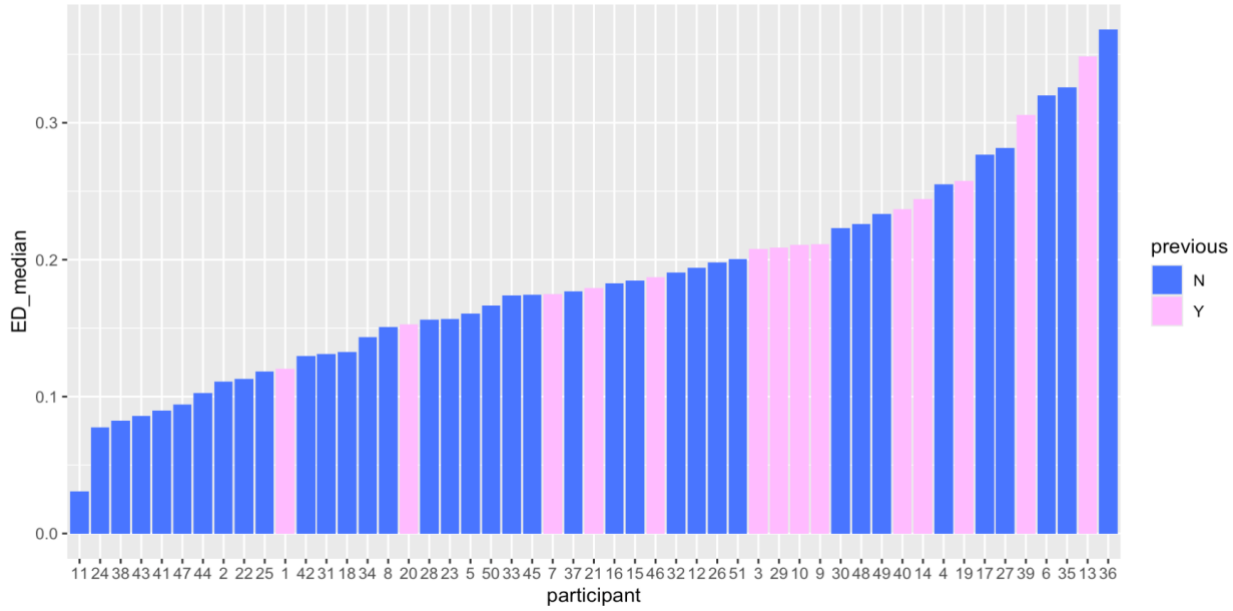


Figure 2.16: Individual ED-median values for **[a]/[ɒ]$_A$** calculated with non-Bark-converted data and coded by PREVIOUS.

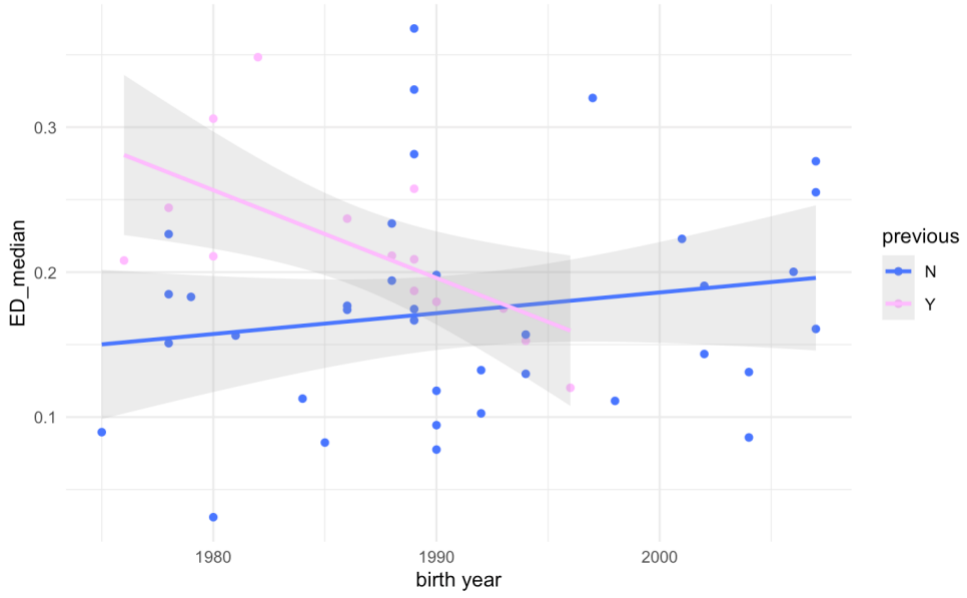


Figure 2.17: Individual ED-median values for **[a]/[ɒ]$_A$** calculated with non-Bark-converted data and coded by BIRTH YEAR and PREVIOUS, with superimposed regression lines. Each dot represents one participant, the x-axis represents BIRTH YEAR, and the color coding represents PREVIOUS.

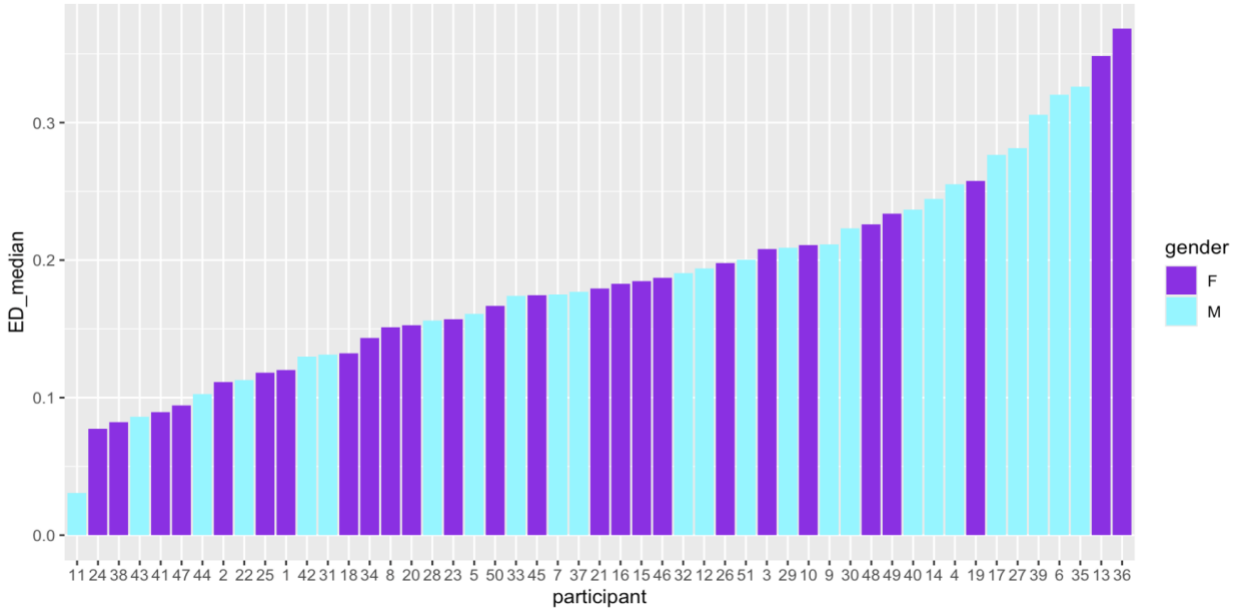


Figure 2.18: Individual ED-median values for **[a]/[ɒ]$_A$** calculated with non-Bark-converted data and coded by GENDER.

In models with Bark-converted data, the results differed slightly. AICc once again preferred the model with PREVIOUS and the interaction of PREVIOUS and BIRTH YEAR. This interaction was once again close to being significant (p=0.088), while GENDER was also close to significant in this model (p=0.078), with men using more dialect-like (i.e., distinct) pronunciations than women. The second-best model according to AICc was the one with GENDER, BIRTH YEAR, EDUCATION LEVEL, and PREVIOUS. In this model, both PREVIOUS (p=0.063) and GENDER (p=0.054) were close to significant, with previous participants and men using more dialect-like pronunciations. BIC preferred the simple model with only GENDER, BIRTH YEAR, and EDUCATION LEVEL, in which GENDER was significant (p=0.046). The results of these three models are presented in **Tables 2.10-2.12**.

Table 2.10: Results of the linear regression model predicting ED-median for [a]/[ɒ]$_A$ calculated with Bark-converted data. The model included PREVIOUS and its interaction with BIRTH YEAR. PREVIOUS was dummy coded (such that the coefficients and p-values for PREVIOUS and BIRTH YEAR do not represent main effects). P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. The adjusted R-squared of the model was 0.136.

Predictor	p-value
GENDER	p=0.078
BIRTH YEAR	p=0.443
EDUCATION LEVEL – non-higher vs. higher	p=0.675
PREVIOUS	p=0.087
BIRTH YEAR – PREVIOUS interaction	p=0.088

Table 2.11: Results of the linear regression model predicting ED-median for [a]/[ɒ]$_A$ calculated with Bark-converted data. The model included PREVIOUS. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. The adjusted R-squared of the model was 0.097.

Predictor	p-value
GENDER	p=0.054
BIRTH YEAR	p=0.965
EDUCATION LEVEL – non-higher vs. higher	p=0.302
PREVIOUS	p=0.063

Table 2.12: Results of the linear regression model predicting ED-median for [a]/[ɒ]$_A$ calculated with Bark-converted data. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. The adjusted R-squared of the model was 0.047.

Predictor	p-value
GENDER	*p=0.046*
BIRTH YEAR	p=0.673
EDUCATION LEVEL – non-higher vs. higher	p=0.589

Figures 2.19-2.21 present the data in comparison with PREVIOUS, the interaction of PREVIOUS and BIRTH YEAR, and GENDER. In accordance with the model results, **Figure 2.21** illustrates a (relatively weak) gender pattern that was not seen in the non-Bark-converted data. This discrepancy suggests that converting data to Bark can affect the results of further analyses in some cases, such that analyses should generally explore discrepancies between results calculated with Bark-converted and non-Bark-converted data.

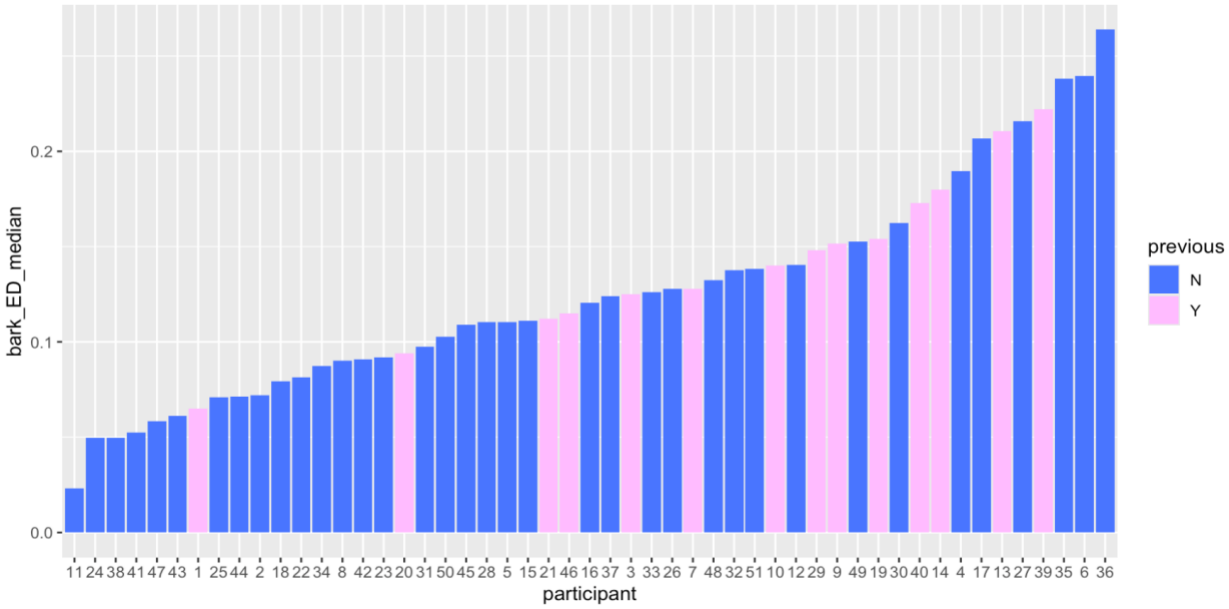


Figure 2.19: Individual ED-median values for **[a]/[ɒ]$_A$** calculated with Bark-converted data and coded by PREVIOUS.

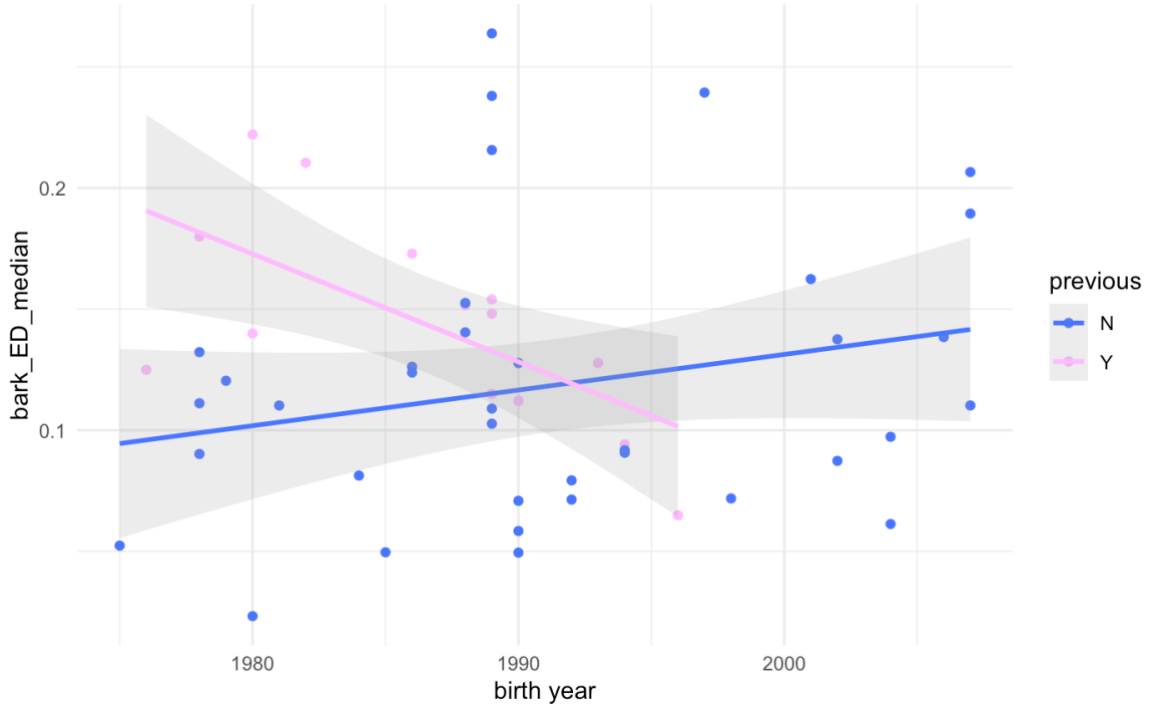


Figure 2.20: Individual ED-median values for **[a]/[ɒ]$_A$** calculated with Bark-converted data and coded by BIRTH YEAR and PREVIOUS, with superimposed regression lines. Each dot represents one participant, the x-axis represents BIRTH YEAR, and the color coding represents PREVIOUS.

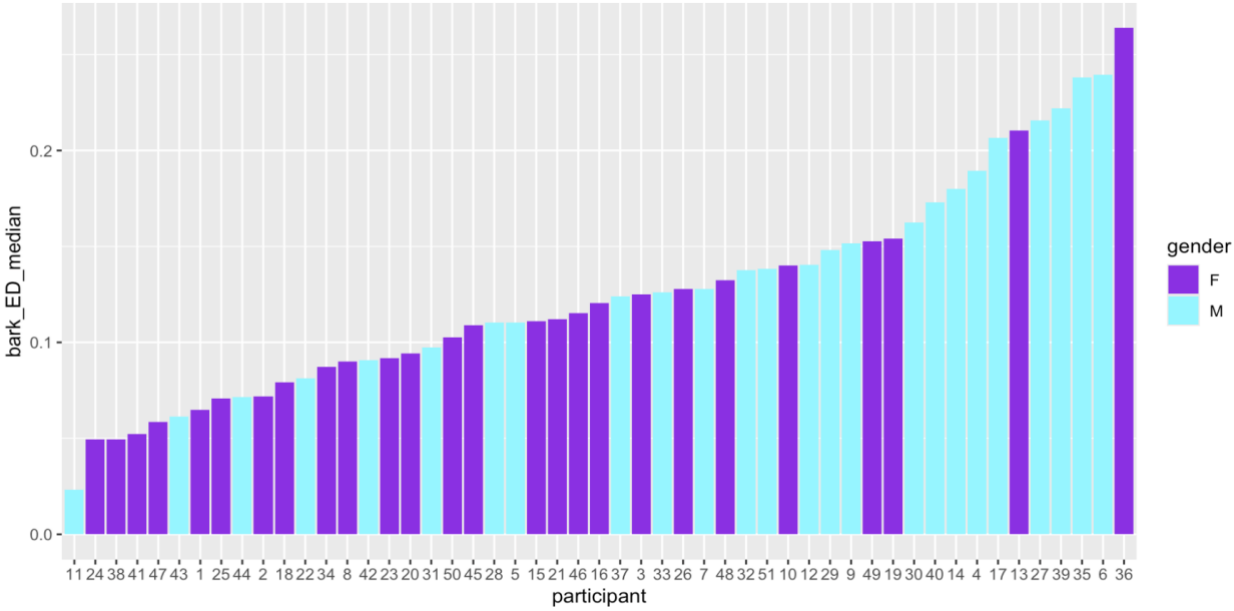


Figure 2.21: Individual ED-median values for **[a]/[ɒ]$_A$** calculated with Bark-converted data and coded by GENDER.

2.5.2 [ɔ]/[ɒ]o

2.5.2.1 Pillai scores

In models predicting Pillai scores for [ɔ]/[ɒ]$_O$, PREVIOUS and its interaction with GENDER turned out to be relevant for predicting the variation in the data. AICc preferred the model that included PREVIOUS and its interaction with GENDER, while BIC preferred the model with PREVIOUS but without the interaction. PREVIOUS is significant in this simpler model (p=0.025), while the interaction of PREVIOUS and GENDER in the more complex model does not reach significance (p=0.114). The results of these models are presented in **Tables 2.13-2.14**. The presented models predicted Pillai scores calculated from non-Bark-converted data, as converting to Bark did not affect Pillai score values (beyond small differences attributable to rounding).

Table 2.13: Results of the linear regression model predicting Pillai scores for [ɔ]/[ɒ]$_O$. The model included PREVIOUS and its interaction with GENDER. Both predictors were dummy coded, such that the coefficient/p-value for GENDER represents the difference between non-previous men and women, while the coefficient/p-value for PREVIOUS represents the difference between previous women and non-previous women. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. The adjusted R-squared of the model was 0.073.

Predictor	**p-value**
GENDER	p=0.455
BIRTH YEAR	p=0.550
EDUCATION LEVEL – non-higher vs. higher	p=0.623
PREVIOUS	*p=0.007*
GENDER – PREVIOUS interaction	p=0.114

Table 2.14: Results of the linear regression model predicting Pillai scores for [ɔ]/[ɒ]$_O$. The model included PREVIOUS. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. The adjusted R-squared of the model was 0.041.

Predictor	**p-value**
GENDER	p=0.873
BIRTH YEAR	p=0.836
EDUCATION LEVEL – non-higher vs. higher	p=0.600
PREVIOUS	*p=0.025*

Figure 2.22 presents individual Pillai scores coded in accordance with PREVIOUS and overlaid with merger/distinction thresholds calculated using the method of Stanley and Sneller (2023), while **Figure 2.23** presents the data coded by GENDER. It can be seen in **Figure 2.22** that previous participants had lower Pillai scores overall, representing less dialect-like pronunciations. **Figure 2.24** visualizes the interaction between PREVIOUS and GENDER. Although this interaction did not reach statistical significance, it is apparent in the graphs that the previous (i.e., repeat) participants show a gender pattern that the new participants do not. The lack of an overall gender pattern (among all the 2025 participants), as shown in **Figure 2.23**, represents a major departure from the results of the 2022 analysis.

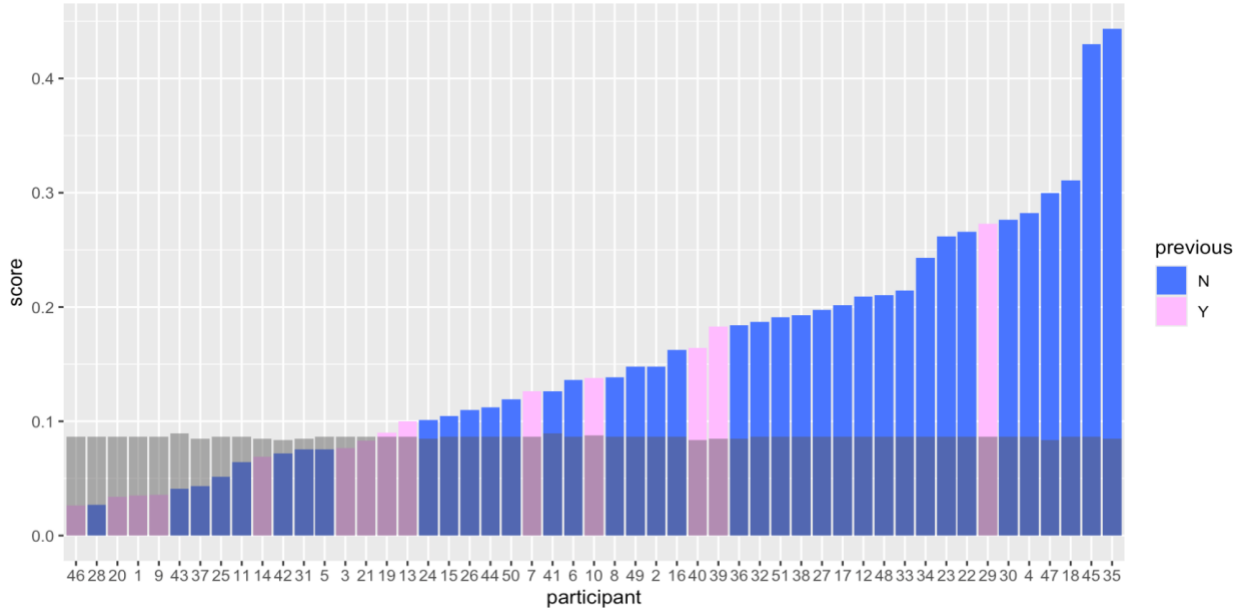


Figure 2.22: Individual Pillai scores for **[ɔ]/[ɒ]o** with superimposed merger/distinction thresholds, coded in accordance with the PREVIOUS predictor.

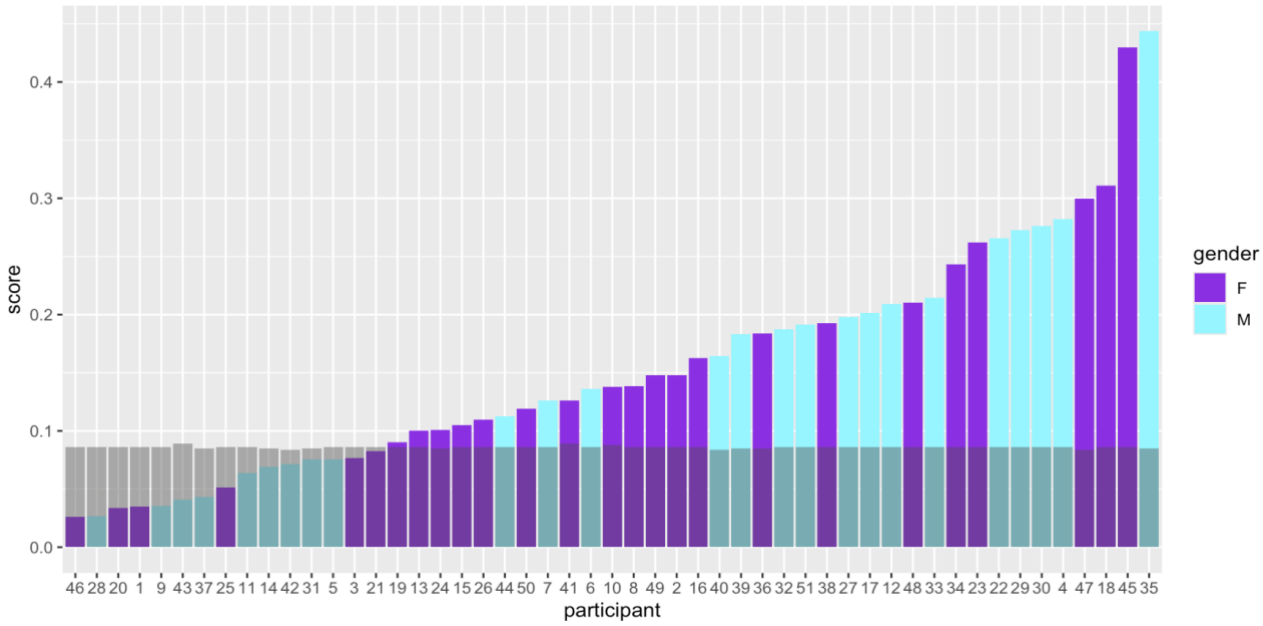


Figure 2.23: Individual Pillai scores by GENDER for **[ɔ]/[ɒ]**$_O$ with superimposed merger/distinction thresholds.

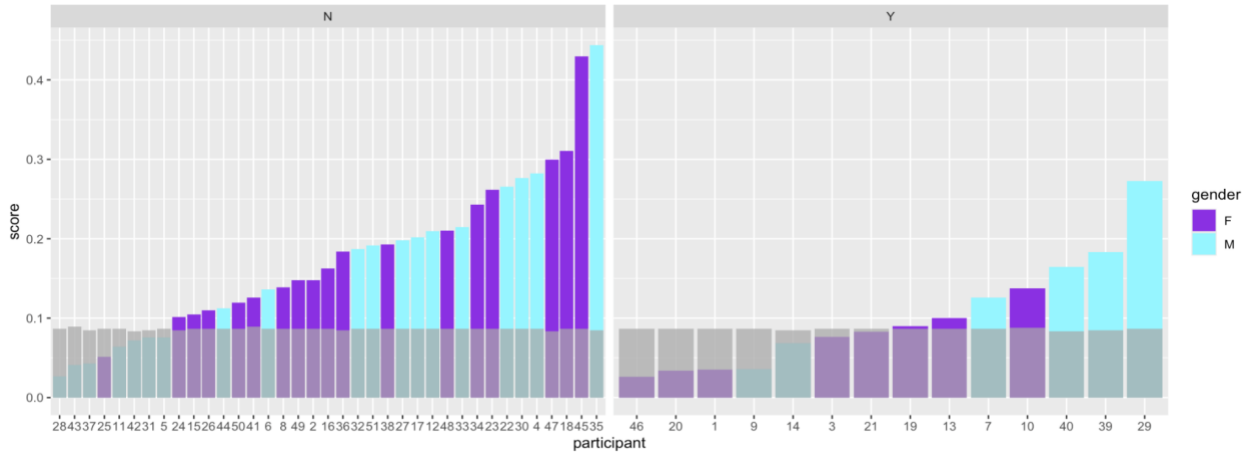


Figure 2.24: Pillai scores of new (left) and previous (right) participants for the comparison of [ɔ]/[ɒ]$_O$ in the 2025 data set. A gender pattern can be seen among the previous participants, but not among the new participants.

2.5.2.2 ED-median

In models predicting ED-median for [ɔ]/[ɒ]$_O$, the interactions of both PREVIOUS and GENDER and PREVIOUS and BIRTH YEAR were found to predict variation in the data. The model that included these interactions was preferred by both AICc and BIC, and its results are presented in **Table 2.15**. **Table 2.16** presents an additional model which included PREVIOUS but no interactions, in order to demonstrate that neither GENDER, nor BIRTH YEAR, nor PREVIOUS showed

significant main effects. The models predicted ED-median calculated from non-Bark-converted data, as converting to Bark did not affect the results in terms of which predictors were significant.

Table 2.15: Results of the linear regression model predicting ED-median for [ɔ]/[ɒ]$_O$. The model included PREVIOUS and its interactions with GENDER and BIRTH YEAR. As the categorical predictors were dummy coded, the individual coefficients/p-values for the predictors involved in the interactions do not represent main effects. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. The adjusted R-squared of the model was 0.140.

Predictor	**p-value**
GENDER	p=0.167
BIRTH YEAR	*p=0.025*
EDUCATION LEVEL – non-higher vs. higher	p=0.613
PREVIOUS	*p=0.031*
GENDER – PREVIOUS interaction	*p=0.008*
BIRTH YEAR – PREVIOUS interaction	*p=0.031*

Table 2.16: Results of the linear regression model predicting ED-median for [ɔ]/[ɒ]$_O$. The model included PREVIOUS. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. The adjusted R-squared of the model was -0.061.

Predictor	**p-value**
GENDER	p=0.718
BIRTH YEAR	p=0.882
EDUCATION LEVEL – non-higher vs. higher	p=0.432
PREVIOUS	p=0.752

Figure 2.25 visualizes the interaction of PREVIOUS and GENDER, showing (just as in the case of Pillai scores), that a gender pattern is present among previous participants but not new participants. **Figure 2.26** visualizes the interaction of PREVIOUS and BIRTH YEAR, showing different relationships between BIRTH YEAR and PREVIOUS among previous and new participants. **Figure 2.27** shows individual ED-median values in relation to GENDER, while **Figures 2.28**-**2.29** show the data in relation to PREVIOUS and BIRTH YEAR respectively. As in the case of Pillai scores, the lack of a gender pattern represents a departure from the results of the 2022 analysis.

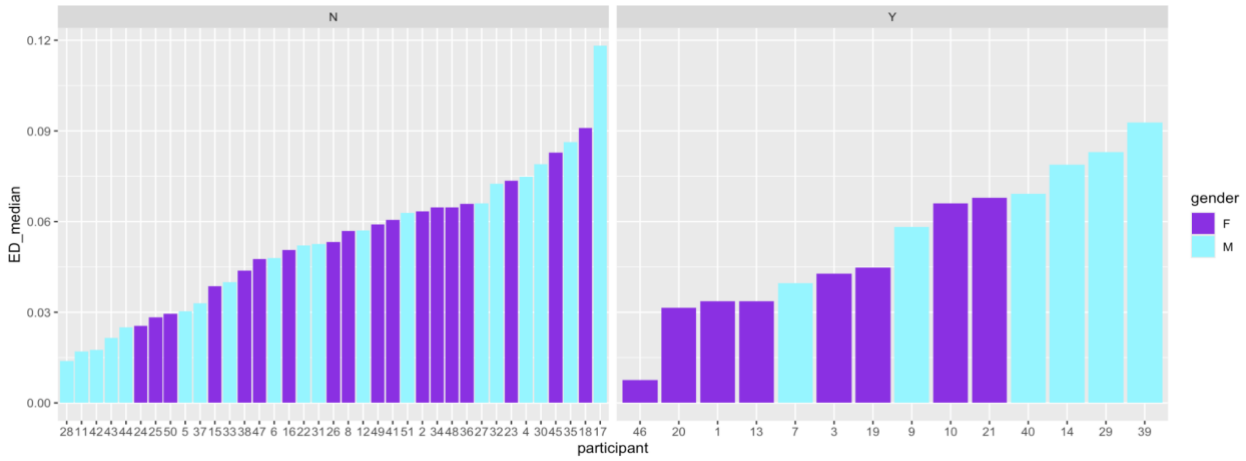


Figure 2.25: ED-median measurements of new (left) and previous (right) participants for the comparison of **[ɔ]/[ɒ]o** in the 2025 data set. A gender pattern can be seen among the previous participants, but not among the new participants.

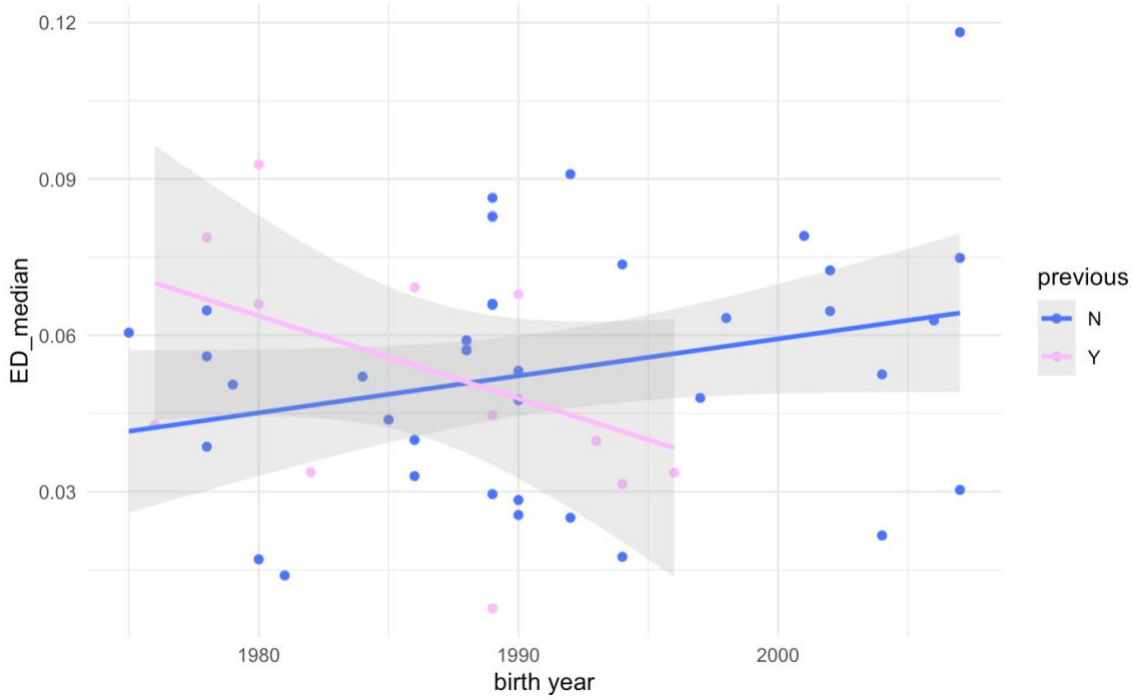


Figure 2.26: Individual ED-median values for **[ɔ]/[ɒ]o** coded by BIRTH YEAR and PREVIOUS, with superimposed regression lines. Each dot represents one participant, the x-axis represents BIRTH YEAR, and the color coding represents PREVIOUS.

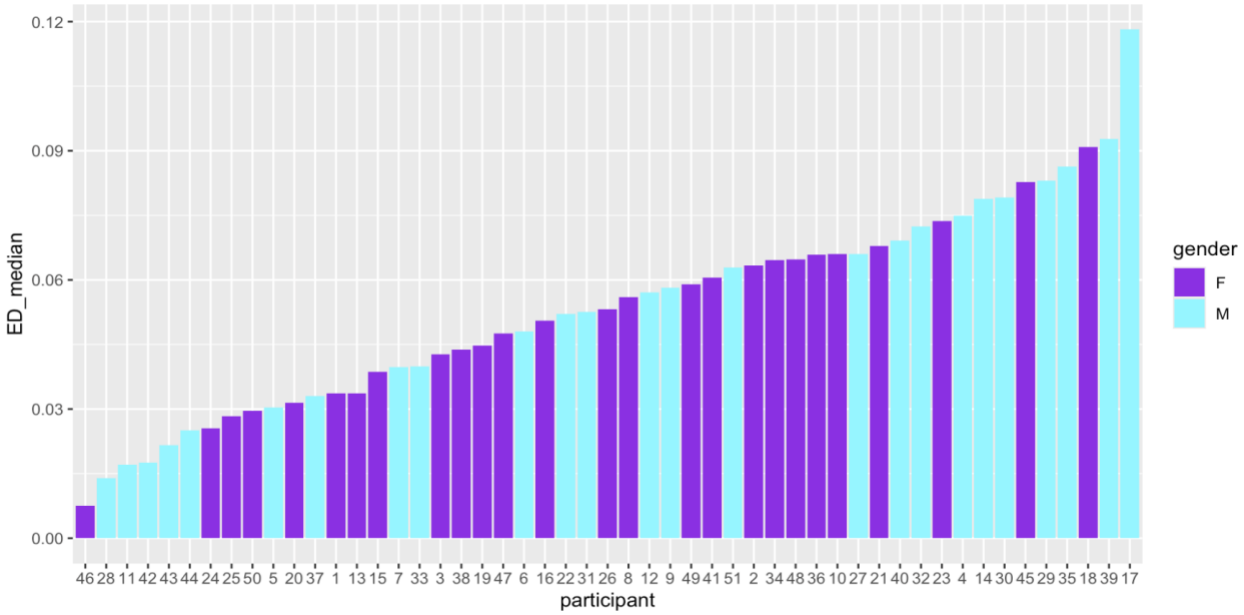


Figure 2.27: Individual ED-median values by GENDER for **[ɔ]/[ɒ]$_O$**.

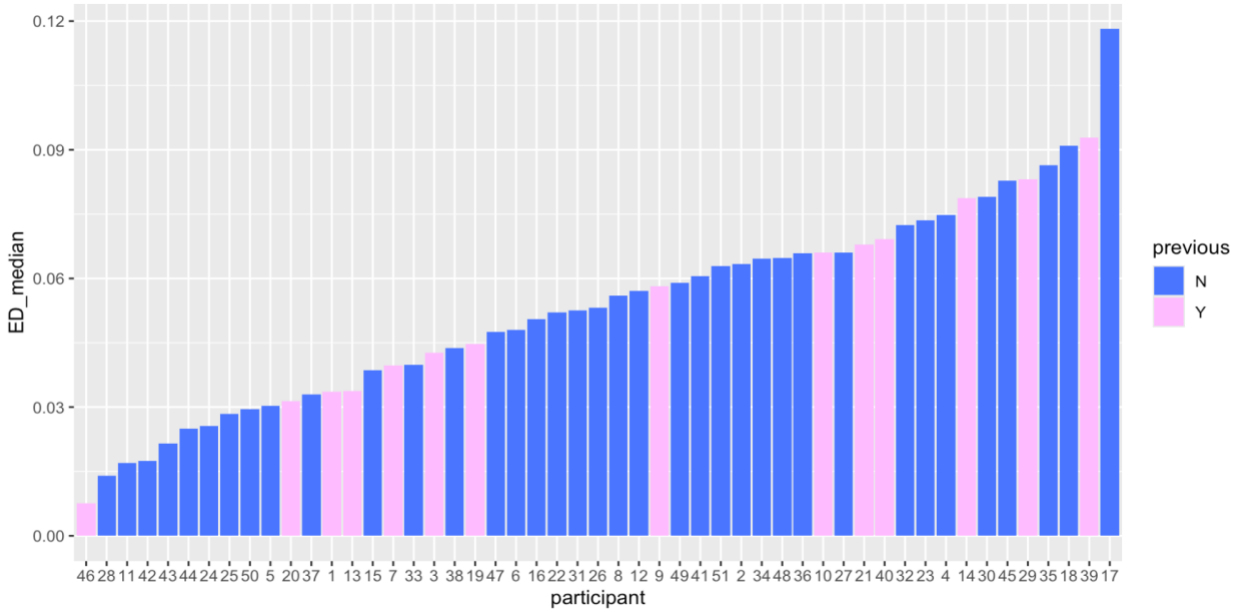


Figure 2.28: Individual ED-median values for **[ɔ]/[ɒ]$_O$** coded in accordance with PREVIOUS.

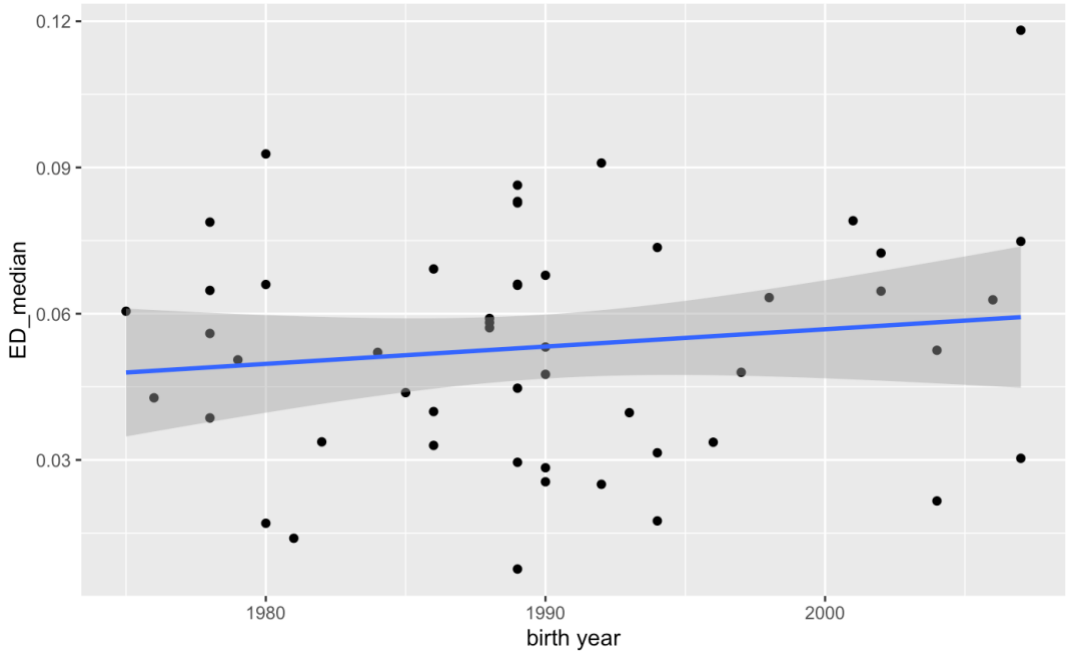


Figure 2.29: Individual ED-median values by BIRTH YEAR for **[ɔ]/[ɒ]$_O$** with superimposed regression line.

2.5.3 [ɒ]$_A$/[ɒ]$_O$

2.5.3.1 Pillai scores

In models predicting Pillai scores for [ɒ]$_A$/[ɒ]$_O$, GENDER was found to be a significant predictor, with men having smaller Pillai scores (more dialect-like pronunciation) than women. This is consistent with the results of the 2022 analysis. BIC preferred the simple model with GENDER, BIRTH YEAR, and EDUCATION LEVEL, while AICc preferred a model that also included the interaction of EDUCATION LEVEL and GENDER, though this interaction did not reach significance (p=0.084). **Tables 2.17-2.18** present the results of these two models. The presented models predicted Pillai scores calculated from non-Bark-converted data, as converting to Bark did not affect Pillai score values (beyond small differences attributable to rounding).

Table 2.17: Results of the linear regression model predicting Pillai scores for [ɒ]$_A$/[ɒ]$_O$. The model included the interaction of GENDER and EDUCATION LEVEL. Both predictors were dummy coded, such that the coefficient/p-value for GENDER represents the difference between men and women with higher education, while the coefficient/p-value for EDUCATION LEVEL represents the difference between women with and without higher education. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. The adjusted R-squared of the model was 0.394.

Predictor	**p-value**

GENDER	*p < 0.001*
BIRTH YEAR	p=0.486
EDUCATION LEVEL – non-higher vs. higher	p=0.843
GENDER – EDUCATION LEVEL interaction	p=0.084

Table 2.18: Results of the linear regression model predicting Pillai scores for $[ɒ]_A/[ɒ]_O$. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. The adjusted R-squared of the model was 0.367.

Predictor	**p-value**
GENDER	*p < 0.001*
BIRTH YEAR	p=0.854
EDUCATION LEVEL – non-higher vs. higher	p=0.114

Figure 2.30 shows individual Pillai scores in relation to GENDER with overlaid merger/distinction thresholds calculated using the method of Stanley and Sneller (2023). **Figure 2.31** depicts the interaction of GENDER and EDUCATION LEVEL, showing that there was a difference between the education groups (people with higher education having more dialect-like pronunciations) among men but not among women.

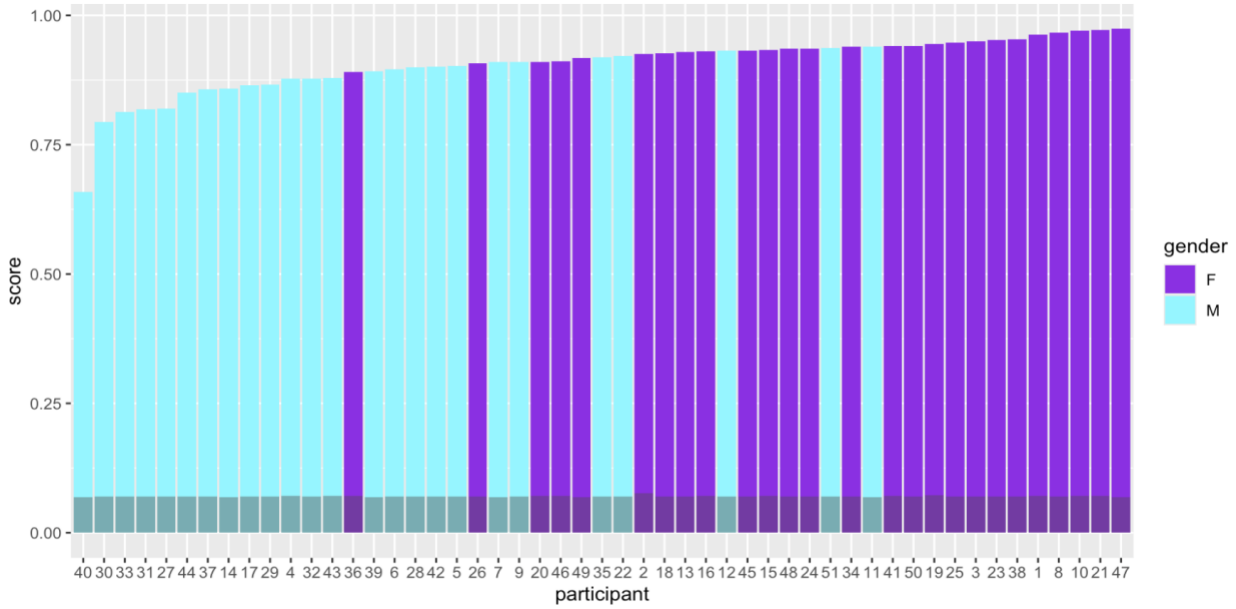


Figure 2.30: Individual Pillai scores by GENDER for **$[ɒ]_A/[ɒ]_O$** with superimposed merger/distinction thresholds.

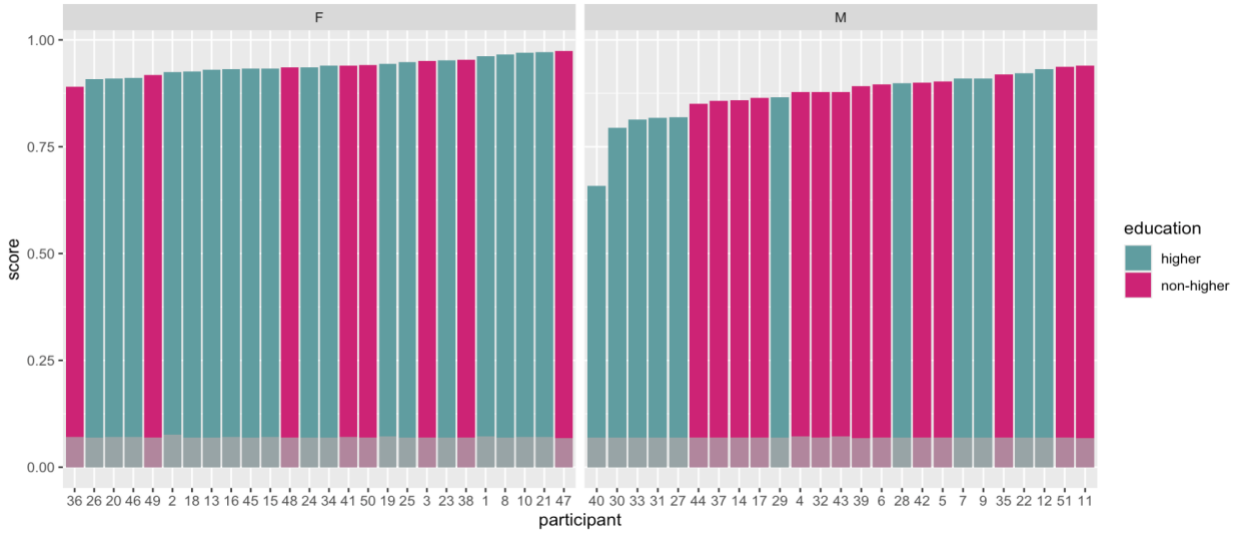


Figure 2.31: Individual Pillai scores for **[ɒ]$_A$/[ɒ]$_O$** grouped by GENDER and EDUCATION LEVEL with superimposed merger/distinction thresholds.

2.5.3.2 ED-median

In models predicting ED-median for [ɒ]$_A$/[ɒ]$_O$, GENDER and PREVIOUS were found to be significant predictors, with men using more dialect-like pronunciations and previous participants using less dialect-like pronunciations. The results of the analyses using non-Bark-converted and Bark-converted data were only slightly different from one another. In the analysis using non-Bark-converted data, PREVIOUS was not quite significant (p=0.064), and BIC preferred the model with only GENDER, BIRTH YEAR, and EDUCATION LEVEL, while AICc preferred the model with PREVIOUS. In the analysis using Bark-converted data, both GENDER and PREVIOUS were significant, though the p-value for PREVIOUS (p=0.040) was similar to the value in the model with non-Bark-converted data. EDUCATION LEVEL is also somewhat close to being significant in both of these models (with people with higher education using more dialect-like pronunciations), similarly to the simpler model for Pillai scores presented in the previous section. The results of these models are presented in **Tables 2.19-2.20**.

Table 2.19: Results of the linear regression model predicting ED-median for [ɒ]$_A$/[ɒ]$_O$ calculated with non-Bark-converted data. The model included PREVIOUS. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. The adjusted R-squared of the model was 0.658.

Predictor	**p-value**

GENDER	*p<0.001*
BIRTH YEAR	p=0.722
EDUCATION LEVEL – non-higher vs. higher	p=0.128
PREVIOUS	p=0.064

Table 2.20: Results of the linear regression model predicting ED-median for [ɒ]$_A$/[ɒ]$_O$ calculated with Bark-converted data. The model included PREVIOUS. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. The adjusted R-squared of the model was 0.596.

Predictor	**p-value**
GENDER	*p<0.001*
BIRTH YEAR	p=0.724
EDUCATION LEVEL – non-higher vs. higher	p=0.084
PREVIOUS	*p=0.040*

Figures 2.32-2.35 present individual ED-median values coded by GENDER and PREVIOUS and calculated with both non-Bark-converted and Bark-converted data.

The very stark gender pattern for ED-median for [ɒ]$_A$/[ɒ]$_O$ is consistent with the 2022 analysis. The extreme similarity of this pattern across the two data sets raises questions about its explanation. Since the Pillai scores for this vowel pair in the 2025 data set show very little overlap between the two categories for any speaker, and because these Pillai scores are likely more precise/accurate than those calculated in 2022 (see Section 2.6.2 below), it would seem that speakers generally maintain a clear difference between [ɒ]$_A$ and [ɒ]$_O$. Thus, despite the very clear pattern of gender-based variation in the ED-median values, it is questionable whether any speaker (at least, in the 2025 sample) can be said to be approximating the dialect vowel system with respect to the relationship between these two vowels. Rather, the stark gender difference in the ED-median measurements may have a more global explanation. Specifically, if men generally have a more compressed vowel space (even after accounting for vowel normalization), which seems to be the case based on **Figures 2.10** and **2.11** above, then their [ɒ]$_A$/[ɒ]$_O$ categories would automatically be closer together than those of women. This also aligns with more general cross-linguistic findings – for example, Diehl et al. (1996: 187) note that female speakers' vowels generally show greater "between-category dispersion," possibly in connection with the

need to balance vowel identifiability with their voices' higher f0. Thus, the stark gender pattern in the ED-median values for these categories may have an explanation that is related to properties of the entire vowel space and to general differences between men's and women's vowel spaces, rather than being necessarily indicative of dialectal pronunciations of the given vowels.

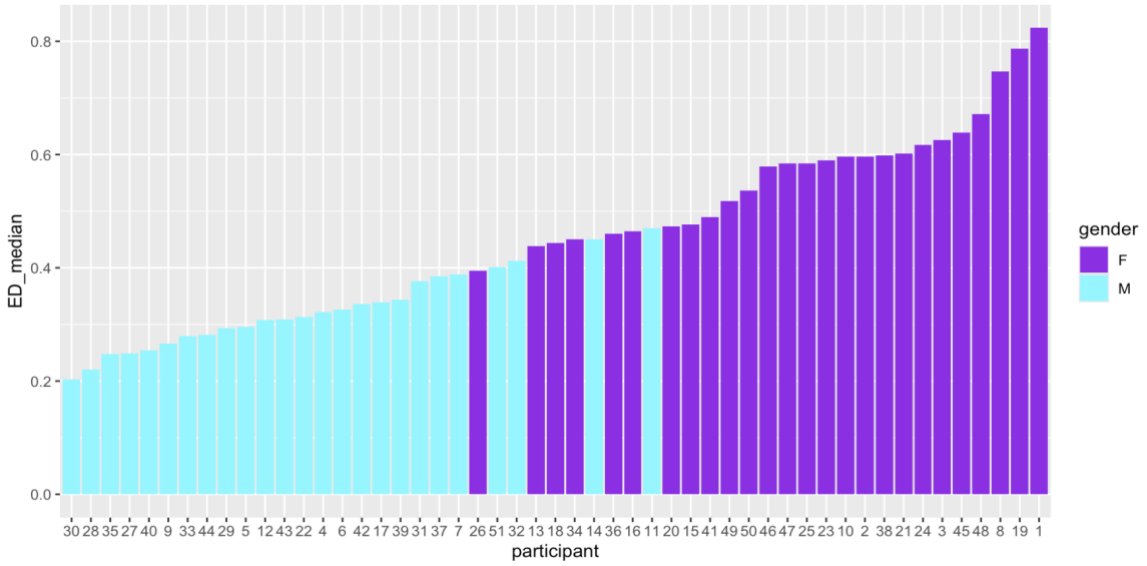


Figure 2.32: Individual ED-median values for $[ɒ]_A/[ɒ]_O$ calculated with non-Bark-converted data and coded by GENDER.

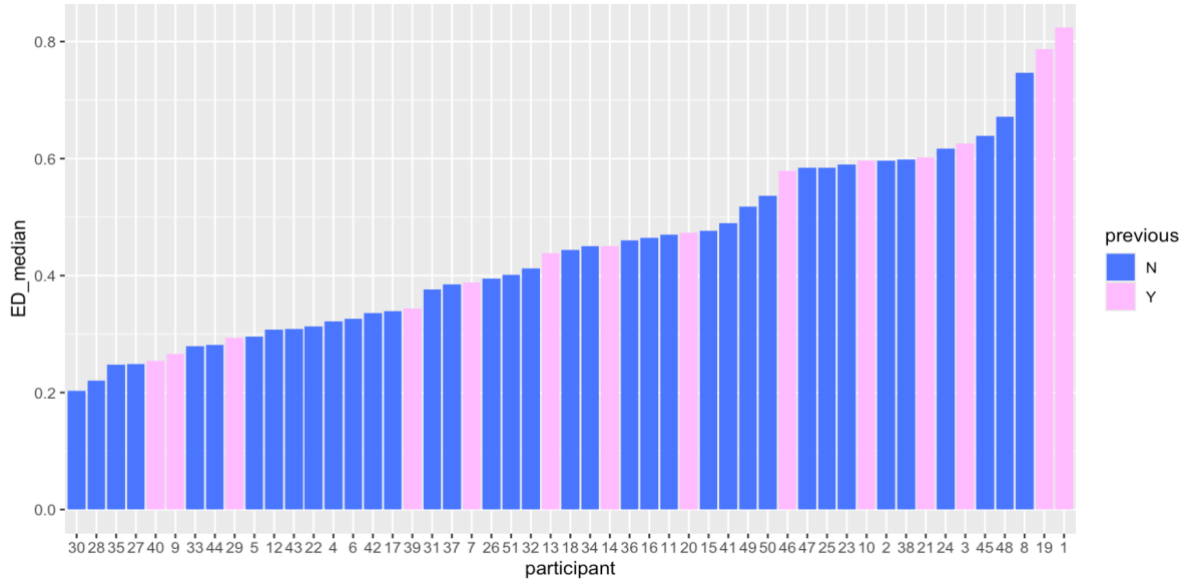


Figure 2.33: Individual ED-median values for **[ɒ]$_A$/[ɒ]$_O$** calculated with non-Bark-converted data and coded by PREVIOUS.

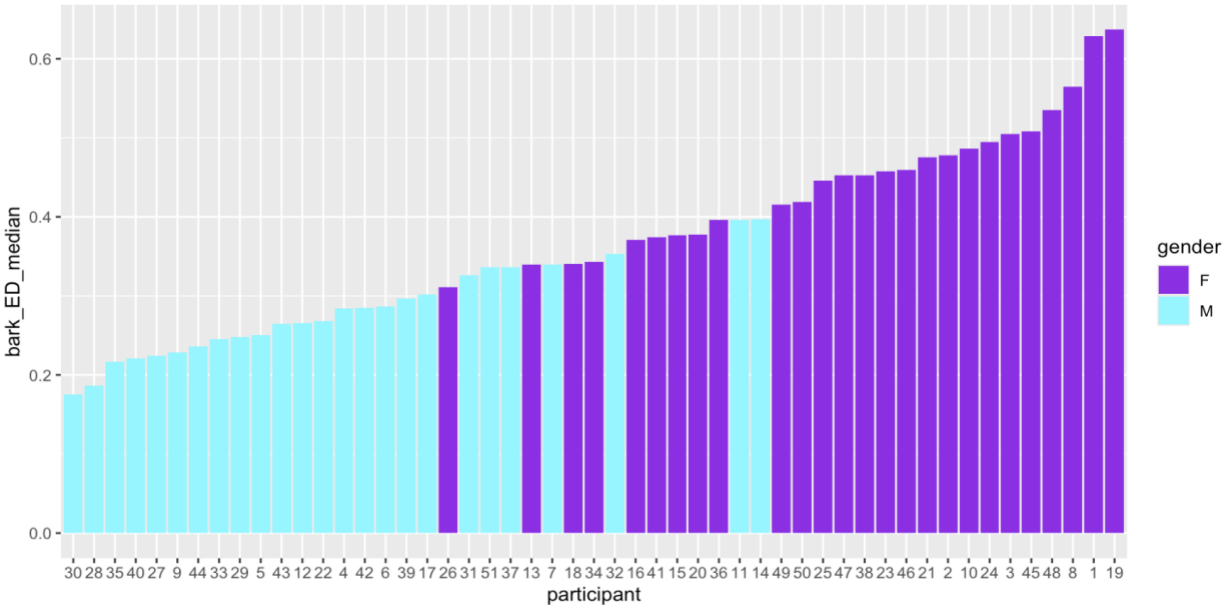


Figure 2.34: Individual ED-median values for **[ɒ]$_A$/[ɒ]$_O$** calculated with Bark-converted data and coded by GENDER.

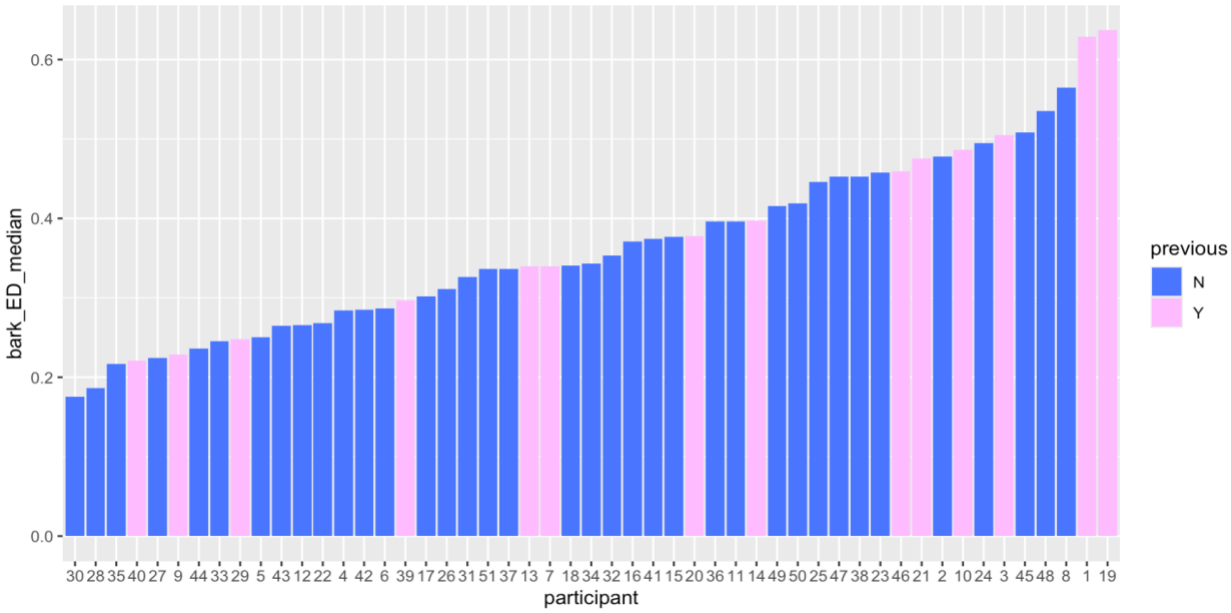


Figure 2.35: Individual ED-median values for **[ɒ]$_A$/[ɒ]$_O$** calculated with Bark-converted data and coded by PREVIOUS.

2.6 Comparison and discussion of 2022 and 2025 production data

As noted in Section 2.5, the main discrepancy between the 2022 and 2025 results was that gender-based variation (with men using more dialect-like vowels than women) was found for the vowel pairs [a]/[ɒ]$_A$ and [ɒ]$_A$/[ɒ]$_O$ in both studies, while such variation was found for [ɔ]/[ɒ]$_O$ only in the 2022 data. As the Figures in Section 2.5 show, the acoustic differences that participants produced for [a]/[ɒ]$_A$ were generally larger than those produced for [ɔ]/[ɒ]$_O$, which could be related not only to the social salience of [a]/[ɒ]$_A$ and its apparently more consistent use in the expression of (gender-based) social meaning, but also to phonological aspects of this vowel pair that will be discussed in Chapter 3.

Section 2.6.1 below compares the 2022 and 2025 results of those participants who participated in both studies; this comparison suggests that the discrepancy in the gender pattern for [ɔ]/[ɒ]$_O$ is a result of participant sampling issues, as a gender pattern is observed for these repeat participants in both data sets, but not for the 2025 participant sample as a whole. With consideration for such effects of participant sampling on the results of the analysis, Section 2.6.1 also discusses the advantages and disadvantages of different participant sampling methods in sociolinguistic research, devoting particular attention to the feasibility of different sampling methods in small communities and those where it may be difficult to elicit colloquial speech.

Comparison of the two data sets also reveals differences in the magnitude and range of Pillai scores and ED-median values for [ɒ]$_A$/[ɒ]$_O$ and [ɔ]/[ɒ]$_O$ between the two data sets. As Section 2.6.2 discusses, this is likely due to differing token sample sizes and differing recording quality (resulting from different recording equipment) between the two studies. This explanation aligns with previous research about the potential effects of these factors.

As Section 2.6.3 discusses, a comparison of words where the vowel [æ] was expected in both studies suggests that participants may have spoken in a more formal or monitored manner in the 2025 study, even when considering only the word-list data from the two studies. This comparison demonstrates that it is important to consider potential effects of elicitation style that are more subtle than those generally expected for different speech styles (e.g. word-list vs. spontaneous speech data).

Overall, though the results of the two studies had many similarities in terms of their findings related to patterns of social variation in the maintenance of dialect vowels, comparison of the results raises questions about important methodological issues in sociolinguistic research.

The results also set the stage for comparison with social variation in the perception of the given vowels, and for investigation of individual-level perception-production links.

2.6.1 Differences in patterns of social variation – sampling issues

As noted above, the main difference between the 2022 and 2025 production results was that gender-based variation was found for the [ɔ]/[ɒ]$_O$ vowel pair in 2022 but not 2025. As discussed in Section 2.5.2, the models predicting Pillai scores and ED-median for this comparison indicated that the interaction between GENDER and PREVIOUS was relevant for explaining variation in the outcome variables. Specifically, the models indicated that men had more dialect-like vowels than women among the previous (i.e., repeat) participants, but not among the new participants. **Figures 2.24**-**2.25** above show these differing patterns among the new and repeat participants in the 2025 data set, while **Figures 2.36**-**2.37** below show that the same type of overall gender pattern can be seen among the repeat participants in both 2022 and 2025. Although the repeat participants are not ordered in the exact same manner across the two data sets, it seems reasonable that the discrepancies between some individuals' productions across the two data sets could have been introduced by factors such as differing recording quality (see Section 2.6.2), differing elicitation methods (see Section 2.6.3), and by the fact that each data set represents speakers' productions on a single day, with three intervening years between them. The relative stability of the pattern found among the repeat participants across multiple data sets in combination with the discrepancy between the new and repeat participants in the 2025 data set suggests that the difference in gender variation across the two data sets is due to problems with participant sampling rather than problems with the formant measurement procedure. Accordingly, this section will discuss various issues related to participant sampling in sociolinguistic research.

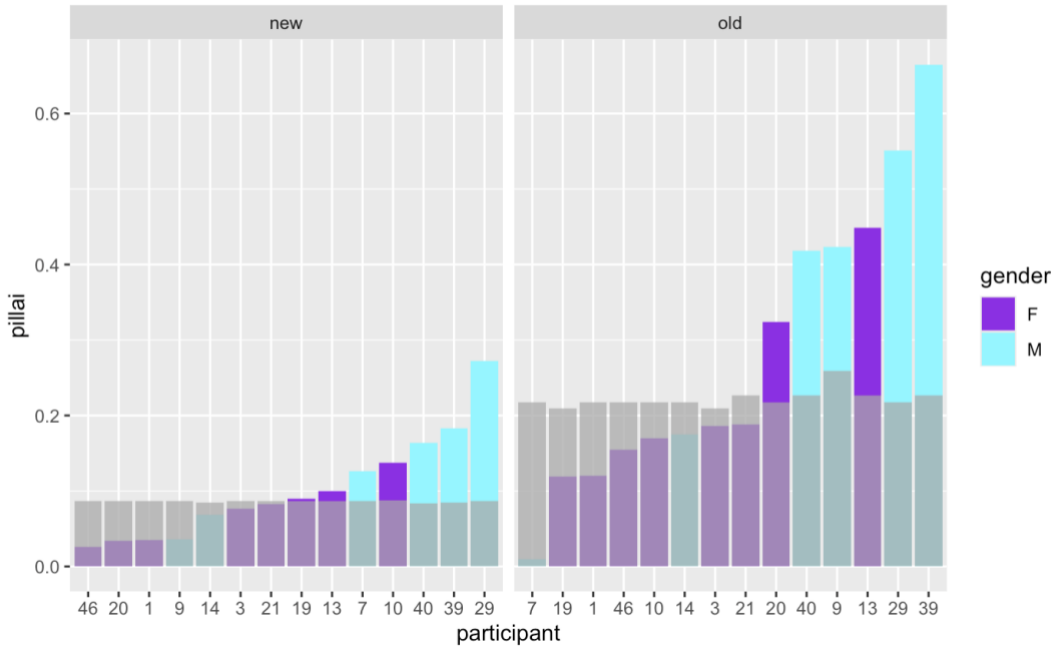


Figure 2.36: Pillai scores of previous (repeat) participants for [ɔ]/[ɒ]$_O$ in the 2025 (left) and 2022 (right) data sets. A similar gender pattern can be seen among these participants across both data sets. Pillai scores are smaller overall in the 2025 data set, likely due to larger sample sizes. Accordingly, although the overlaid merger/distinction thresholds calculated with the method of Stanley and Sneller (2023) differ across the two graphs, a similar number of participants fall above and below the thresholds in both data sets.

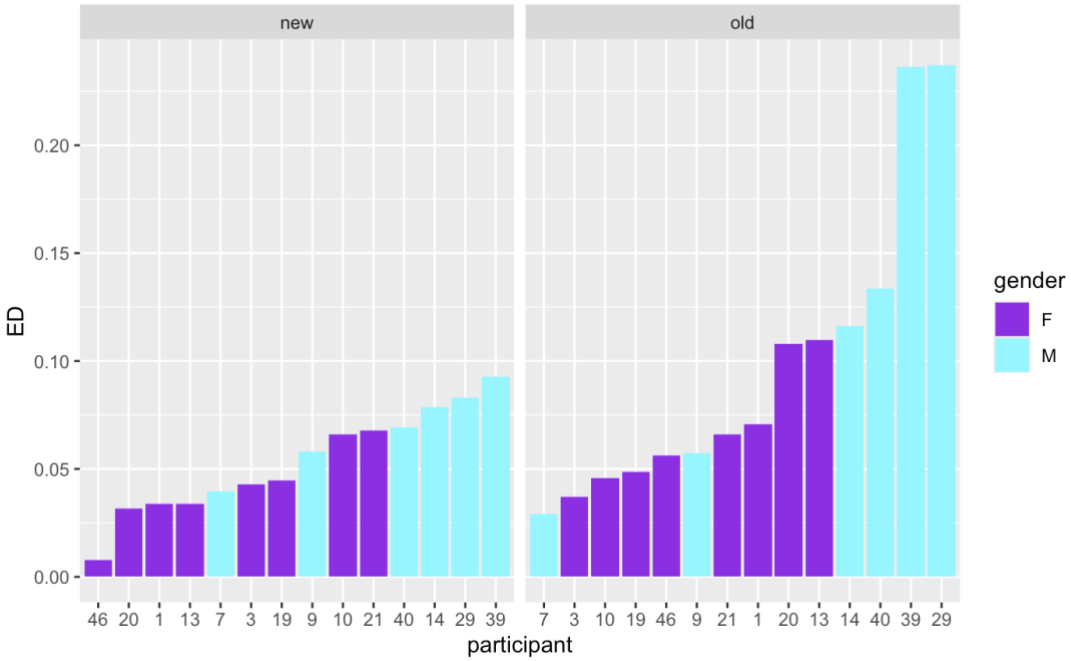


Figure 2.37: ED-median measurements of previous (repeat) participants for the comparison of [ɔ]/[ɒ]$_O$ in the 2025 (left) and 2022 (right) data sets. A similar gender pattern can be seen among these participants across both data sets.

Various types of sociolinguistic research have employed different participant sampling methods in accordance with their goals. As noted in Section 2.3.3.1, traditional dialectology generally seeks out the most conservative speakers in a community (often "nonmobile, older, rural males" [Chambers & Trudgill 1998]). On the other hand, variationist sociolinguistics generally seeks a random sample, which can be considered more representative of the speech community as a whole (Labov 2001: 38-39). Labov (2001: 38-40) stresses that a group of participants who exist within the same social network and/or who have personal connections with the researchers cannot be an accurate representation of an urban speech community, while also acknowledging that some researchers (e.g., Haeri 1996; Milroy and Milroy 1978) have deliberately used non-random sampling methods either out of necessity or in accordance with their research goals (e.g., examining relationships between linguistic variation and social network connections). Labov (2001: 38-40) also notes that both types of sampling methods (random sampling vs. making use of social networks and personal connections) have advantages and disadvantages, as non-random sampling methods are more likely to allow researchers to elicit vernacular speech and overcome the observer's paradox.

Though Labov (2001: 38) argues that one should strive for a random sample in which "each one of several million speakers has an equal chance of being selected," it is debatable whether this is even a realistic ideal to strive toward. Beyond the general bias toward the inclusion of speakers who are willing to participate in a research study (whose speech could differ in some systematic way from the speech of people who are not willing to participate), the distribution of speakers in a community may not always be such that every type of speaker will be included in a modestly sized random sample. In comparing sampling methods used in traditional dialectology and modern sociolinguistics, Maguire (2007: 169-170) explains that "traditional dialect speakers," who are generally not very numerous in a given community, could be entirely missed in a sample of several dozen speakers, if that sample is "not designed to uncover rarer speech types."

While Labov (2001: 38-39) states that a reliable sample of a "very large" city can be achieved with less than one hundred speakers, which would make a sample of 30-50 speakers seem adequate for a small city like Gavaṛ (whose population was 17,741 according to the 2022 census), the sampling methods used in both the 2022 and 2025 studies nevertheless have many shortcomings. As participants in the 2022 study were all previously acquainted with the

interviewer and generally also had immediate and/or extended family members in the sample, we can make limited inferences about the speech community in general based on this non-random sample. The 2025 sample was also not random, in that those participants who were not previously acquainted with the local research assistant were nevertheless recruited through her social networks. As both studies were conducted in the summer, when many men from Gavaṛ go to Russia for work, it is also possible that men in the samples (who maintained some dialect vowels more than the women in the samples) are not representative of men from Gavaṛ as a whole, as the speech of men who do and do not remain in Gavaṛ year-round could differ in some systematic way. Despite the shortcomings of these sampling methods, it is possible that using random sampling in Gavaṛ would have prevented access to colloquial speech to such an extent as to make the results of the studies completely uninformative. Given Armenia's diglossic environment (see Chapter 1) and given that many participants of the 2022 study stated that they change their speech when conversing with people from other parts of Armenia, it is possible that interacting with an unfamiliar local interviewer (whom participants may have initially assumed to be non-local, due to knowing that I am non-local) would have led to increased style-shifting toward SEA. Thus, the sampling method's contribution to obtaining more natural speech (especially in the 2022 study) may have outweighed the limitation of the lack of generalizability of the results. It is also possible that obtaining an ideal random sample, in which participants had no direct contact with one another (Labov 2010: 190-191), would have been prohibitively difficult in Gavaṛ, as the city's small population would have made it difficult to find enough participants who had no connection with anyone else in the sample.[18]

The differences between the 2022 and 2025 results could have several explanations related to sampling issues. For example, the 2022 sample could be less representative of the population than the 2025 sample due to the 2022 participants potentially being more tightly connected within a social network. It is possible that gender-based variation of the [ɔ]/[ɒ]o vowel pair is characteristic of some subsets of the population but not others, and accordingly, the 2022 sample could have represented one such subset of the population. On the other hand, it is possible that one or both samples were simply too small to give an accurate picture of the population, regardless of the sampling method. Given the degree of linguistic heterogeneity that

[18] People in Gavaṛ (not study participants) whom I met in completely different contexts often turned out to know one another.

exists in Armenia, and the corresponding degree of variation in vowels that was observed in the data, a much larger number of participants might have been required in order to make correlational variationist sociolinguistic methods maximally informative. In general, despite the shortcomings of the sampling methods used in the 2022 and 2025 studies, it is clear that one must carefully weigh the advantages and disadvantages of different sampling methods when working with small populations and in situations where it may be difficult to access colloquial speech. In the 2022 sociolinguistic interviews in particular, though the direct social-network sampling (i.e., all participants previously acquainted with research assistant) could explain the finding of gender-based variation in [ɔ]/[ɒ]$_{O}$ that may not generalize to the broader population, there is also evidence that the sampling and elicitation methods contributed considerably to the elicitation of more natural speech (see further discussion in Section 2.6.3 below).

2.6.2 Potential effects of token sample sizes and recording quality

Figures 2.36 and **2.37** above show that while the gender pattern among the repeat participants for [ɔ]/[ɒ]$_{O}$ was similar in 2022 and 2025, the actual Pillai score and ED-median values were generally much smaller in the 2025 data set. Though one might suspect that these differences indicate actual differences in the maintenance of the dialect vowel [ɒ]$_{O}$ across the two data sets, the height of the merger/distinction thresholds (calculated with the method of Stanley & Sneller [2023]) on the two graphs in **Figure 2.36** call this interpretation into question. A substantial difference in the height of the thresholds can be seen in accordance with the sample sizes of the two data sets – specifically, the thresholds are lower for the 2025 data set, in which the median sample size for this comparison was 63 tokens, as opposed to 25 tokens in 2022. Importantly, the thresholds appear to fall at a similar place on the two graphs relative to the range of Pillai scores; that is, a similar number of repeat participants fall above and below the thresholds in both data sets. Accordingly, it is likely that the difference in the range of Pillai scores found across the two studies reflects exactly the issue that Stanley and Sneller's (2023) threshold method is intended to address – namely, that Pillai scores are generally larger when sample sizes are smaller, even if the underlying population is truly merged. Interestingly, as **Figure 2.38** below shows, there was not a substantial difference in the range of Pillai scores for [a]/[ɒ]$_{A}$ in the 2025 and 2022 data. For this vowel pair, the median per-speaker sample sizes were 79 and 38 in the 2025 and 2022 studies respectively. Therefore, although there was a difference in sample sizes between the two studies for both vowel pairs, the sample size in the

2022 study for [ɔ]/[ɒ]$_O$ was substantially smaller than the one for [a]/[ɒ]$_A$. The fact that there is such a discrepancy between 2022 and 2025 Pillai scores for [ɔ]/[ɒ]$_O$ but not for [a]/[ɒ]$_A$ suggests that there perhaps exists some sample size threshold below which we can expect significant inflation of Pillai scores, but above which Pillai scores may be more stable across different sample sizes. In other words, the difference in sample sizes for [ɔ]/[ɒ]$_O$ may have led to a large difference in Pillai scores because the 2022 sample size of approximately 25 tokens per speaker was simply very low, while the 2022 sample size of approximately 38 tokens per speaker for [a]/[ɒ]$_A$ may not have been low enough to cause noticeable inflation of Pillai scores.

The large discrepancy in ED-median values for [ɔ]/[ɒ]$_O$ across the two data sets (see **Figure 2.37**) can perhaps be explained in a similar manner. Like with Pillai scores, a similar discrepancy was not observed for [a]/[ɒ]$_A$. It is possible that the small sample sizes for [ɔ]/[ɒ]$_O$ in 2022 led to more uncertainty in the location of the centers of different groups of tokens, making the groups seem farther apart than they would with increased sample sizes that would make the true central tendency of each group more clear. The effects of differing sample sizes across the two studies were likely compounded by differences in recording quality caused by using different microphones and recorders in the two studies. The use of different microphones and recording devices has been found to affect formant measurements (e.g., Freeman & De Decker 2021), as has recording quality (e.g., Rathcke et al. 2017). The quality of recordings in the 2022 study was generally lower, which likely led to increased uncertainty in the location of formants and thus increased uncertainty in measures of their central tendency. Thus, it is possible that both small sample sizes and poorer recording quality contributed to the larger ED-median measurements seen in the 2022 data for [ɔ]/[ɒ]$_O$.

It should be noted that these same issues (sample size and recording quality) likely contributed to the discrepancy between the Pillai scores for [ɒ]$_A$/[ɒ]$_O$ in 2022 vs. 2025. For this vowel pair (see Section 2.5.3 above), Pillai scores were overall higher (indicating less overlap) and had a much smaller range in 2025 than in 2022. This is likely due to the apparent increased precision in measurements of the central tendency of the [ɒ]$_O$ category, thanks to increased sample size and better recording quality. In general, the differences between the two data sets with respect to [ɔ]/[ɒ]$_O$ and [ɒ]$_A$/[ɒ]$_O$ align with the observations of previous authors (e.g., Stanley & Sneller 2023; Freeman & De Decker 2021; Rathcke et al. 2017) in suggesting the

importance of considering factors like token sample size and recording quality when comparing different data sets.

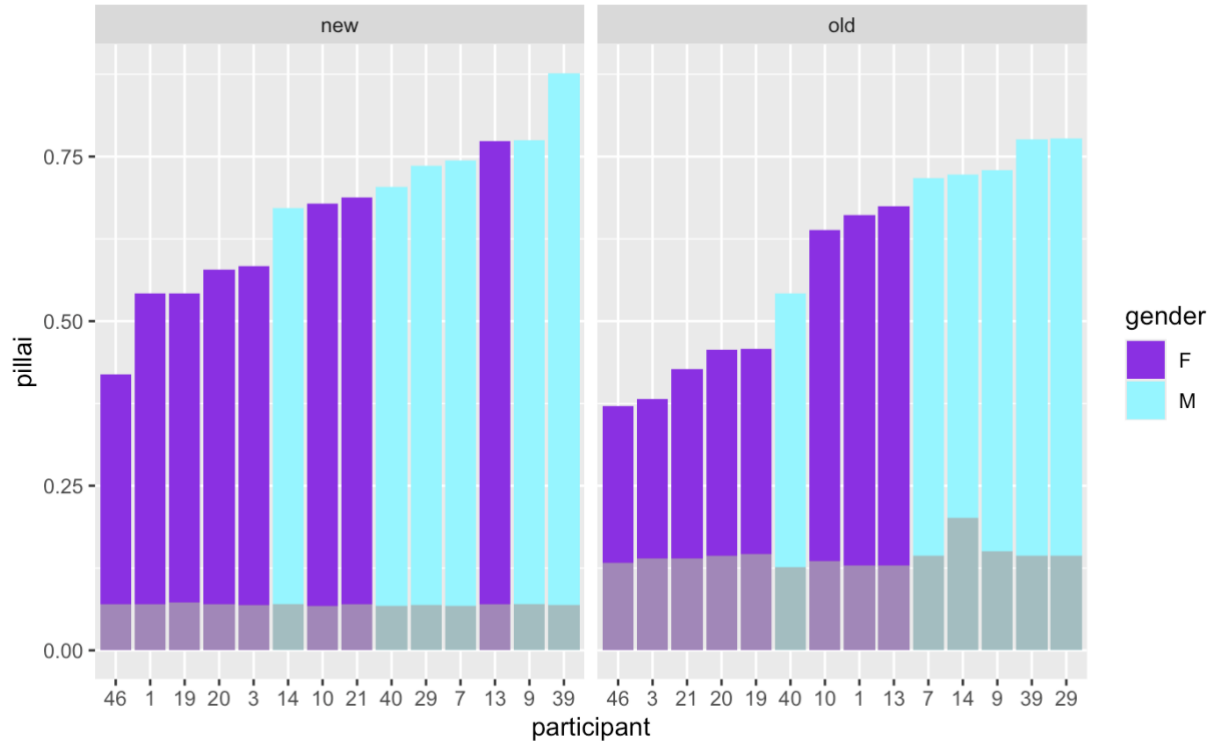


Figure 2.38: Pillai scores of previous (repeat) participants for the comparison of [a]/[ɒ]$_A$ in the 2025 (left) and 2022 (right) data sets, with merger/distinction thresholds calculated using the method of Stanley & Sneller (2023).

2.6.3 Potential effects of elicitation methods

Though the group of words where [æ] is expected was not the main focus of the 2025 study, words of this group were nevertheless included in the word list, and their differing productions in 2022 and 2025 provide insights into potential effects of the differing elicitation methods in the two studies. **Figure 2.39** shows the formant values of each participant's [æ] tokens (red dots) in comparison with those tokens where [a] was expected (black ellipses) in the word list of the 2022 study, while **Figure 2.40** provides analogous individual graphs for the 2025 study. **Figure 2.39** includes only the word-list tokens of [æ] from 2022 (as opposed to both the word and picture tokens), as these tokens are more comparable to the 2025 data (which comes entirely from a word list). As **Figure 2.39** shows, pronunciations of tokens where [æ] was expected were often categorically different from those where [a] was expected in the 2022 word

list. In contrast, **Figure 2.40** shows that such pronunciations were much rarer in 2025, being completely absent for the majority of participants. This discrepancy can be even more clearly seen in the individual data of some repeat participants. For example, Participants 23 and 31 of the 2022 study, who correspond to Participants 29 and 40 of the 2025 study respectively, both had clear groups of tokens with an [æ]-like pronunciations in 2022, but no such tokens in 2025.

The discrepancies between the 2022 and 2025 [æ] data across the two studies indicate that the differing elicitation methods (in addition to different sampling methods – see Section 2.6.1 above) may have affected the degree to which colloquial/dialectal speech was successfully elicited. The 2022 interviews were conducted in quiet rooms either at participants' homes / the homes of their relatives (28 out of 31 interviews), or at my accommodation (3 interviews). Additionally, the 2022 word list was embedded within a larger sociolinguistic interview. By the time participants were reading the 2022 word list, they had already been conversing for approximately 15 minutes and had completed the picture naming task (which one would expect to elicit more [æ]-like pronunciations than the word list due to the lack of potential influence of orthography). In contrast, the majority of participants completed the 2025 study at an unfamiliar location (the Geghark^{h}unik^{h} Regional Library) and recorded the word list immediately after the signing of the informed consent form. Given these differences and the differences in the results, it seems possible that the 2025 study elicited a more formal or monitored style of speech than the 2022 study. It is interesting that the discrepancy between the two data sets is the most apparent for the [æ] data – as discussed in Section 2.3.2.3, the fact that this vowel is both contrastive with [a] (i.e., not phonetically conditioned) and potentially salient may lead to speakers having more conscious control over it than over the other dialect vowels. Though the 2022 and 2025 data showed discrepancies in the gender pattern found in relation to [ɔ]/[ɒ]$_{O}$, the repeat participants generally showed similar patterns across the two studies in relation to the back rounded vowels (see e.g., **Figures 2.36**-**2.37**). Thus, the differing elicitation methods seem to have had the most obvious effects on pronunciations of tokens where [æ] was expected, though less obvious effects on the pronunciation of other vowels cannot be ruled out, especially since some individual participants had substantially different productions across the two studies despite the similar overall patterns at the group level (see, e.g., **Figures 2.36**-**2.37**).

Despite the apparent success of the 2022 interviews in eliciting colloquial/dialectal speech (as evidenced by the seemingly more formal speech obtained in 2025), the methods used

in that study nevertheless have limitations. Although measures were taken to minimize the effects of the "observer's paradox" (e.g., familiar interviewer and interview location), explicitly eliciting language forms is nevertheless not considered an optimal method for uncovering speakers' actual language use (see Milroy & Milroy 2012: 15). While the 2022 data clearly suffered from this issue to some degree, many speakers produced many dialect vocabulary variants during the picture task and also maintained some aspects of the dialect vowel system during the picture and word tasks. This suggests that the unconventional method of explicitly instructing speakers to use their everyday language was likely justified given the local context of relatively high metalinguistic awareness about the existence of different language varieties. It is also possible that some speakers intentionally exaggerated dialect features (like the participant discussed in Section 2.4.3.1), though it is difficult to gauge the extent of such a phenomenon. Overall, while comparison of the 2022 and 2025 results suggests that the 2022 interviews (including the picture and word tasks) may have elicited more natural/colloquial/dialectal speech – especially with respect to the [æ] vowel – the exact influences of the different elicitation methods on the data cannot be fully known. Nevertheless, comparison of the two data sets demonstrates that the general style of elicited speech (e.g., spontaneous speech vs. read speech) is not the only factor that should be considered when comparing different data sets, as more subtle factors (such as the broader context of elicitation) can give rise to systematic differences between data sets that are ostensibly of the same style.

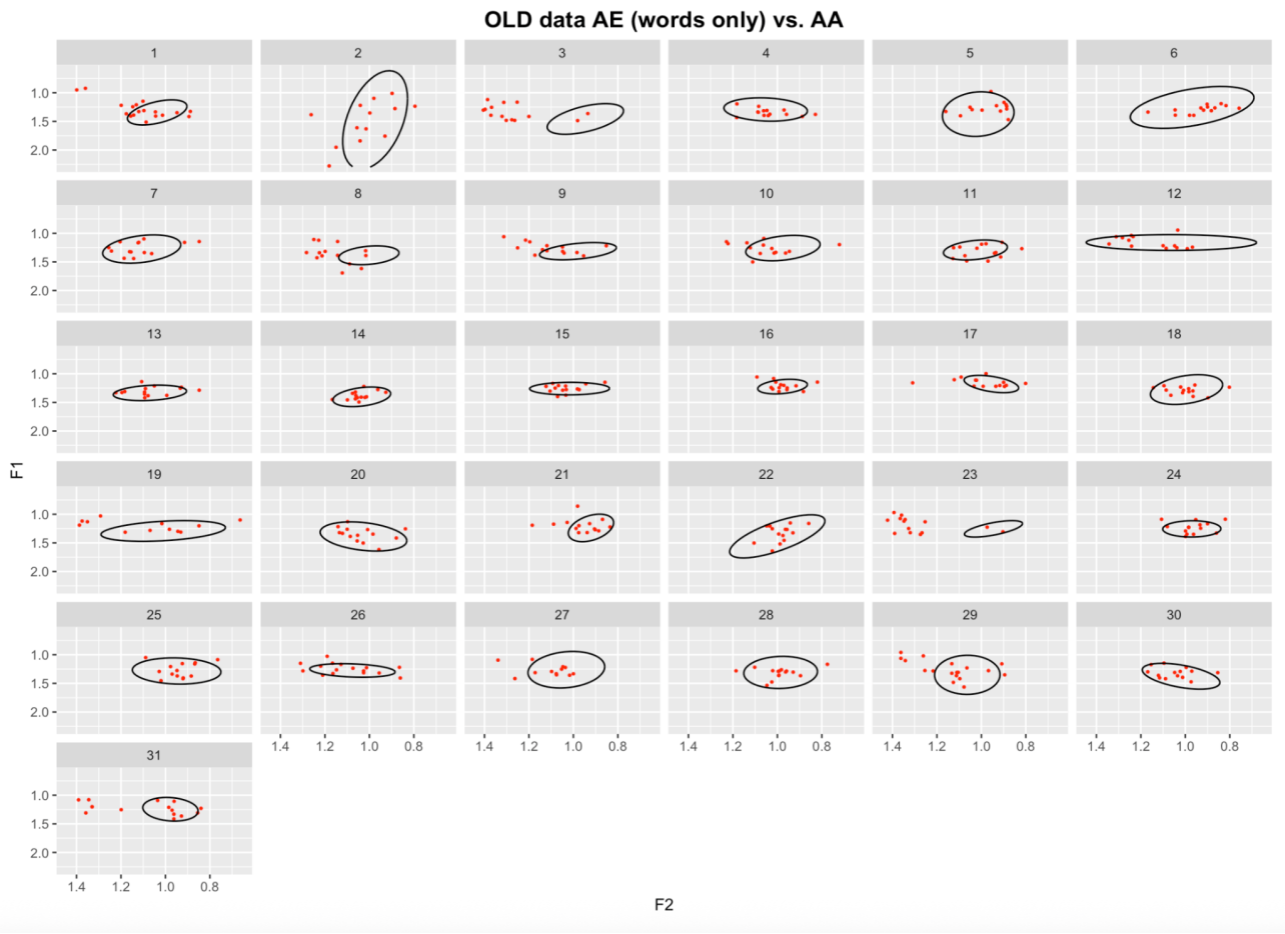


Figure 2.39: Individual formant measurements in normalized units for the group of tokens where [æ] was expected (word list only – individual tokens represented by red dots) and the group of tokens where [a] was expected (black ellipses representing two standard deviations from the mean) in 2022.

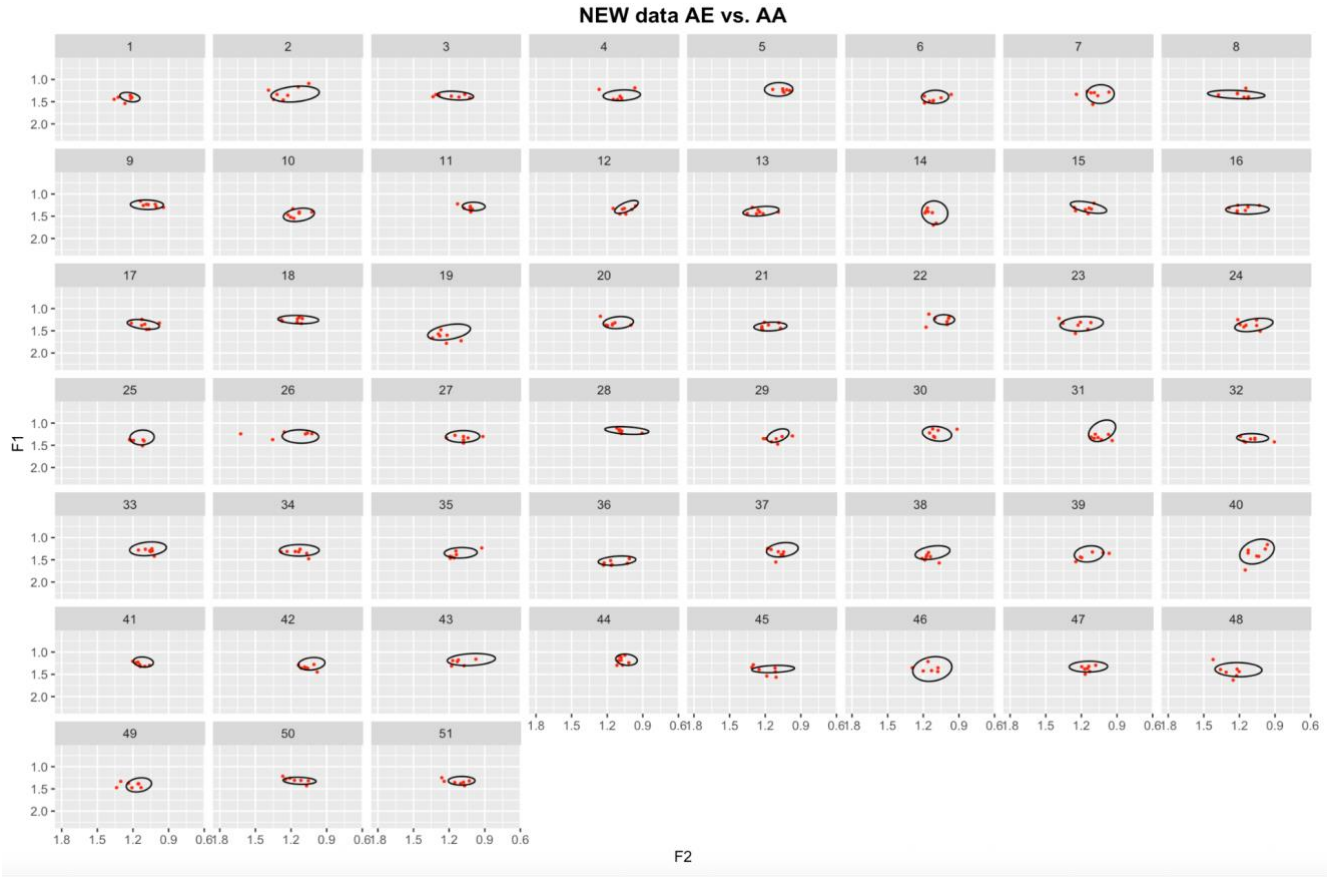


Figure 2.40: Individual formant measurements in normalized units for the group of tokens where [æ] was expected (individual tokens represented by red dots) and the group of tokens where [a] was expected (black ellipses representing two standard deviations from the mean) in 2025.

2.7 General summary of production results

This chapter has presented and compared the results of the 2022 and 2025 production studies, which explored social variation in the maintenance of dialect vowels among Gavaṛ residents. The most consistent result across the two analyses was gender-based variation (with men using more dialect-like vowels than women), which was found for the vowel pairs [a]/[ɒ]$_A$ and [ɒ]$_A$/[ɒ]$_O$ in both studies and for [ɔ]/[ɒ]$_O$ only in the 2022 study. This chapter has also discussed potential explanations for discrepancies between the two data sets, which include participants sampling methods, token sample size and recording quality, and elicitation methods. As noted in Section 2.6, the vowel pair [a]/[ɒ]$_A$ not only showed more consistent patterns of social variation across the two data sets than [ɔ]/[ɒ]$_O$, but also generally showed larger acoustic differences between the two members of the pair, despite being allophonic. As [a]/[ɒ]$_A$ and [ɔ]/[ɒ]$_O$ are the two vowel pairs that were used as stimuli in the perception experiment, the next

chapter (Chapter 3) will motivate the perception experiment by elaborating on certain additional aspects of the production results in relation to these vowels.

Chapter 3 Motivations for Perception Experiment and Overview of Methods and Procedures

3.1 Chapter summary

This chapter presents the research questions and motivations of the two-part perception experiment presented in this dissertation. The experiment was motivated both by aspects of the results of the 2022 study (the results of which are presented in Portugal [2025]), and by questions arising from previous literature. As such, the chapter reviews relevant aspects of the 2022 results, along with relevant literature related to vowel mergers and near-mergers, perception-production relationships, perception of coarticulation and allophonic variation, and effects of social information on speech perception. The chapter then provides a broad overview of the experiment design and implementation, along with an overview of the data collection procedures.

3.2 Perception experiment research questions

The perception study was conducted in 2025 and investigated perception of the vowel pairs [a]/[ɒ]$_{A}$ and [ɔ]/[ɒ]$_{O}$. It addressed four specific questions, which arose both from the results of the 2022 production study and from questions emerging from relevant literature. The 2022 participants' widely varying productions of one allophonic vowel difference ([a]/[ɒ]$_{A}$) raised questions about what it means for vowels to be distinct from one another in a given language variety, and about the role of perceptibility of differences in defining distinctiveness of vowels (see Section 3.3). A review of relevant literature (see Section 3.4) raised questions about differences between different perceptual abilities (operationalized through task type), about the effects of social factors on perception of non-contrastive phonetic detail (e.g., allophony and coarticulatory effects), and about the implications of the 2022 production results for definitions of vowel merger and near-merger.

These broad questions were explored through the four specific research questions of the 2-part perception experiment conducted in 2025:

1) How do residents of Gavaṛ discriminate differences between various vowels of the local dialect's vowel system?

2) How do residents of Gavaṛ categorize such vowels in a task that explicitly elicits metalinguistic judgements about features of the local dialect?

3) How might participants' discrimination of vowel differences in 1) relate to or differ from their categorization of vowels in 2)?

4) How do participants' productions of vowel differences compare with their perceptions, and how might individual participants differ in this regard?

The remainder of this chapter provides an in-depth review of the motivations for these questions, along with a broad overview of the methods and procedures of the perception experiment. More detailed discussion of each research question and of the methods of each part of the study can be found in the subsequent chapters. The first research question is explored in Chapter 4, the second question is explored in Chapter 5, and the third and fourth questions are explored in Chapter 6 (though some comparison of perception and production is integrated into the analyses discussed in Chapters 4 and 5, and some comparison of the results of the two perception tasks is introduced in Chapter 5).

3.3 Motivations from 2022 production results

Several aspects of the 2022 results raised questions about how certain dialect vowels would be perceived, and about how distinctiveness between vowels should be defined with consideration for variation in both perception and production. First, the attempt to calculate merger/distinction thresholds using the method of Stanley and Sneller (2023) highlighted the difficulty of defining what it means for two vowels to be distinct in a given speech community. In the 2022 data, the thresholds calculated for some of the vowel comparisons did not correspond with the observed social variation in any clear way. For example, while **Figure 3.1** below shows that the thresholds for [ɔ]/[ɒ]$_{O}$ seem to divide the unmerged group (mostly men) from the merged group (mostly women), the thresholds for [a]/[ɒ]$_{A}$ and [ɒ]$_{A}$/[ɒ]$_{O}$, as shown in **Figures 3.2** and **3.3**, do not correspond to any critical point in the pattern of variation, as gender patterns are observed despite most speakers' Pillai scores being substantially over the thresholds. (These figures and all other data visualizations in this chapter were created in R [R Core Team 2021] with *ggplot2* [Wickham 2016]).

Thus, it is unclear not only whether these thresholds correspond to what listeners would perceive as distinct, but also whether the thresholds are potentially useful for capturing the moment at which the differences between vowels become socially meaningful. This in turn raises larger questions about what the role of local perception and acoustic measurements should be in determining whether two sounds are distinct in a given language variety, and in defining what it generally means for two vowels to be distinct. For example, if a reliable acoustic difference is found between two sounds, but members of the local speech community do not perceive them as distinct and/or attribute social meaning to their distinction, then which metric should take precedence in the characterization of the language variety's sound system?

Certain characteristics of the vowel system of Gavaṛ's dialect are particularly relevant to an investigation of these questions. Specifically, the comparison of [a]/[ɒ]$_A$ showed results that suggest that consistently measurable differences are not always perceptible or socially meaningful differences. As discussed in Chapter 1, according to Katvalyan (2016: 81-82), [ɒ]$_A$ is an allophonic variant of [a] in Gavaṛ's dialect, which appears word-initially and in the first syllables of words after nasals and voiceless unaspirated consonants except [χ]. As **Figure 3.2** below shows, all speakers in the 2022 study had Pillai scores above the calculated thresholds for this comparison, and examination of speakers' individual vowel plots showed that all speakers' differences were in the expected direction (with the [ɒ]$_A$ cluster being further back and often also higher than [a]). Even so, the differences produced between the two groups by many speakers seemed small enough to potentially be imperceptible. **Figures 3.4** and **3.5** show the individual vowel plots of Speakers 3 and 4, who had the largest and smallest differences between the [a] and [ɒ]$_A$ groups respectively. As the graphs show, in terms of the frontness/backness of the vowels, Speaker 4's difference between the two groups is in the same direction as Speaker 3's much larger difference. **Figure 3.6** shows the plot of Speaker 11, who also has a relatively small difference between the two clusters, which matches the direction of Speaker 3's difference in both backness and height.

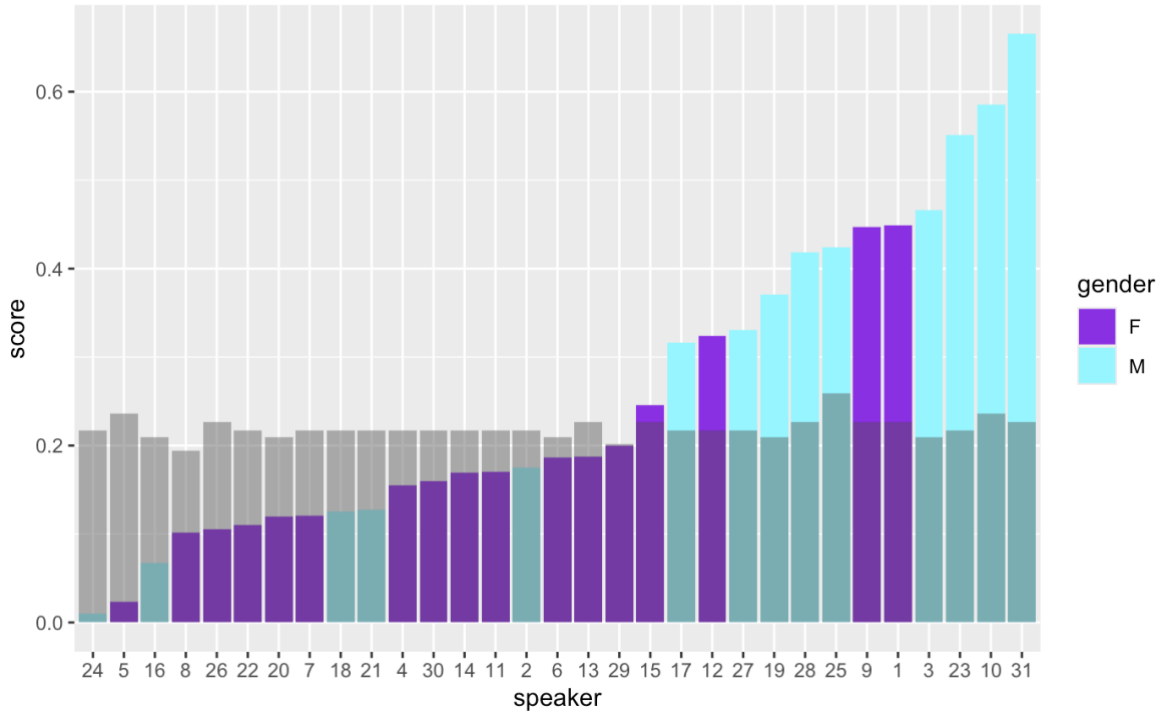


Figure 3.1: Individual Pillai scores (2022 data) by GENDER for **[ɔ]/[ɒ]$_{O}$** with superimposed distinction thresholds.

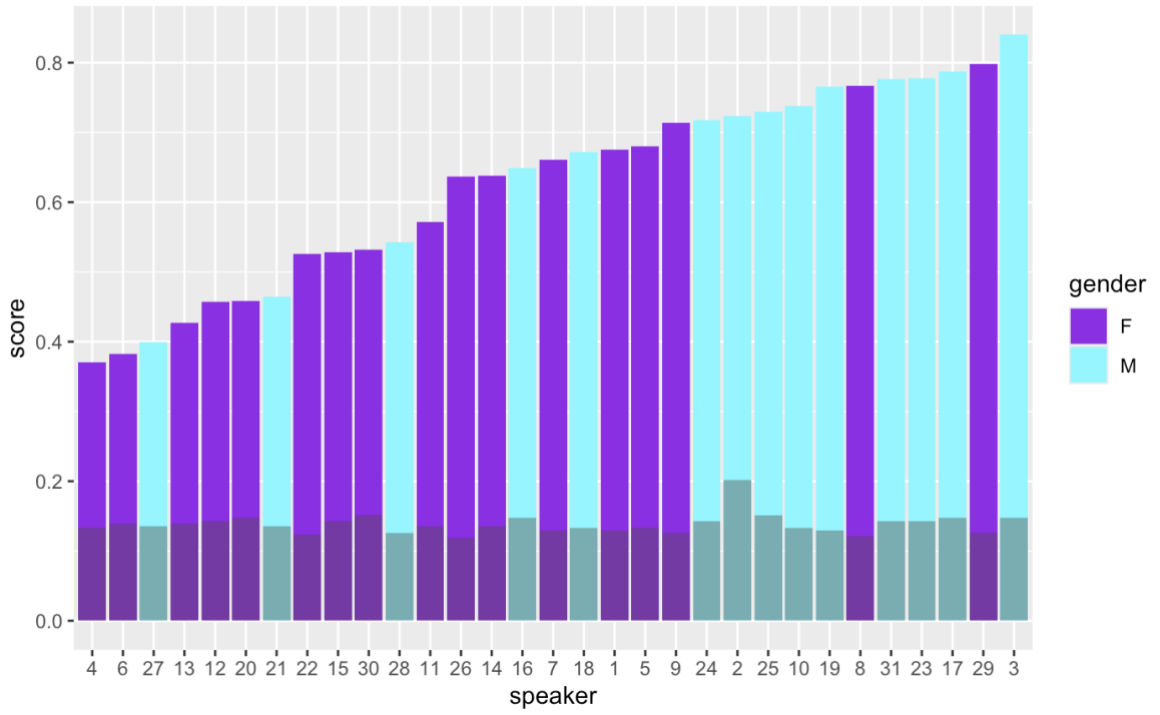


Figure 3.2: Individual Pillai scores (2022 data) by GENDER for **[a]/[ɒ]$_{A}$** with superimposed distinction thresholds.

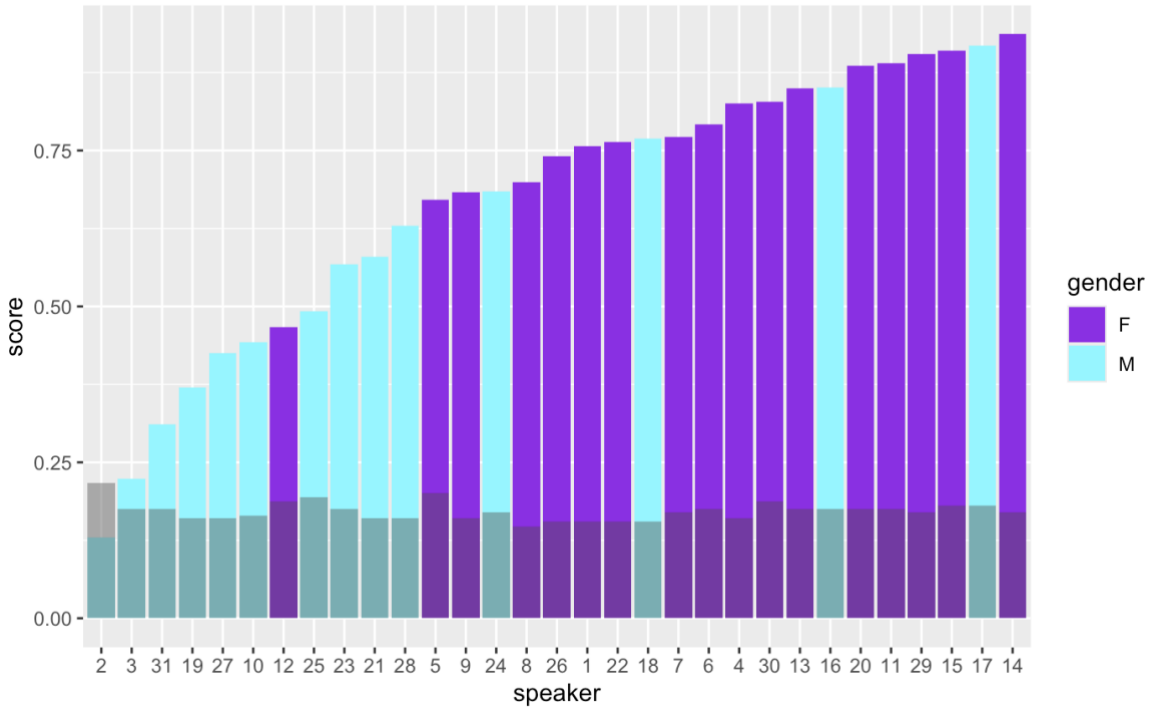


Figure 3.3: Individual Pillai scores (2022 data) by GENDER for **[ɒ]$_A$/[ɒ]$_O$** with superimposed distinction thresholds.

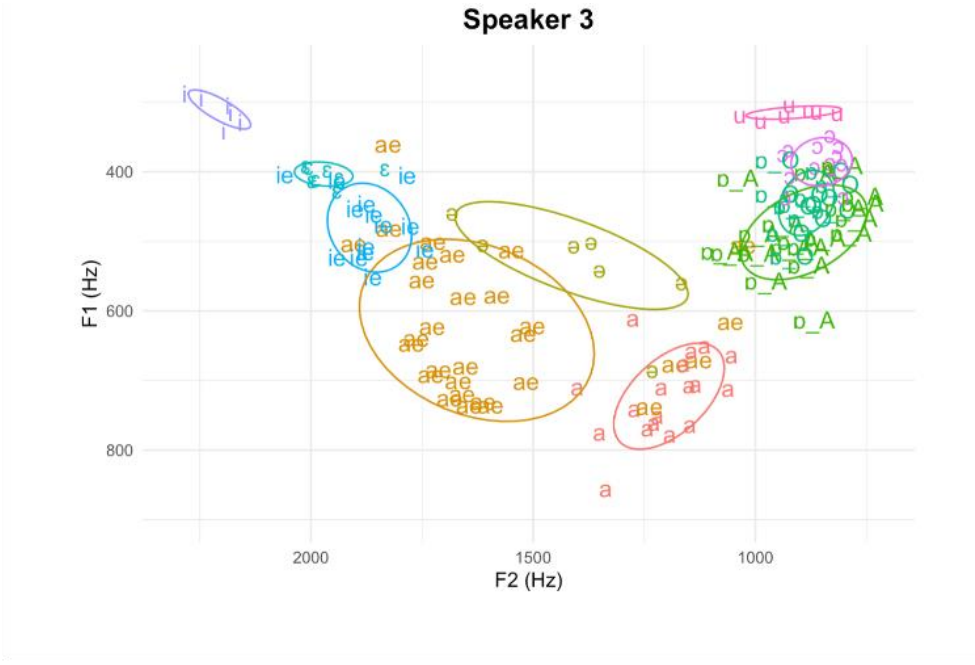


Figure 3.4: Individual vowel plot of Speaker 3 from 2022.

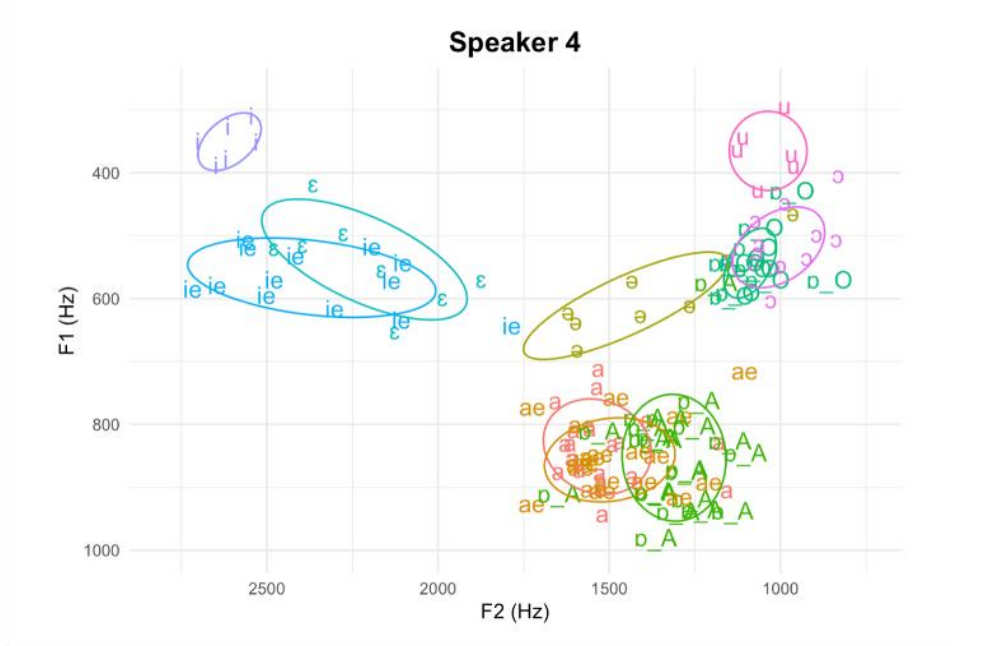


Figure 3.5: Individual vowel plot of Speaker 4 from 2022.

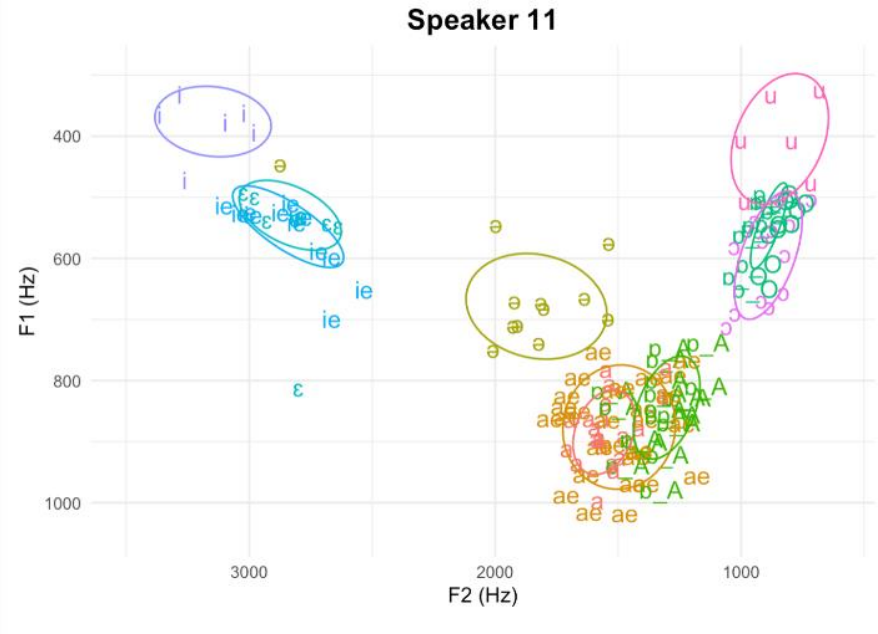


Figure 3.6: Individual vowel plot of Speaker 11 from 2022.

This pattern, in which no speaker shows complete overlap of the two categories despite some of the differences being very small and possibly imperceptible, could be interpreted as a sound change in progress, in which speakers who are more advanced exhibit residual small

differences. However, in the case of these particular vowels, the pattern may also be explained by the potential roots of the dialectal allophony in still-active coarticulatory processes, which would prevent the clusters from ever reaching a point of total overlap. Specifically, Seyfarth and Garellek (2018) found that the vowel [a] in Yerevan Armenian is more front after voiceless aspirated and voiced stops than after plain voiceless stops, while also finding that the voiced series is actually breathy-voiced. Garrett (1998: 17-18) proposed a possible coarticulatory mechanism that could underlie this finding. He writes that the production of breathiness during the release of a stop closure requires a greater drop in transglottal pressure than during the release of a plain voiced stop, and that tongue root advancement (which results in vowel fronting) could be used for this goal.[19] Gavaṛ's dialect also has breathy-voiced stops according to Katvalyan (2016),[20] so it is possible that a similar sort of coarticulatory fronting of [a] after breathy or aspirated sounds lies at the root of the [a]/[ɒ]$_A$ allophony, with this coarticulatory process having become enhanced at some previous point in time. This could explain the small differences between [a] and [ɒ]$_A$ observed for participants like Speaker 4, if we consider that such speakers might be exhibiting the same sort of unenhanced coarticulation as speakers in Yerevan, especially given the increase in contact with Yerevan's colloquial variety over the recent decades. A similar case of enhancement of coarticulation was described by Petersen (2024), who found that a coarticulatory process present in multiple varieties of Danish has become exaggerated and phonologized into prominent allophony in some varieties of Argentine Danish.

There are nevertheless several issues with comparing Seyfarth and Garellek's (2018) findings with the situation of [a]/[ɒ]$_A$ allophony in Gavaṛ. At a glance, based on the previously described environment of the allophony, it might seem that [ɒ]$_A$ appears in the first syllables of words whenever it is not preceded by aspiration or breathiness (including when it is preceded by no consonant). However, there are several problems with this characterization. Firstly, Gavaṛ's dialect is said to have both plain voiced and breathy-voiced stops and affricates in the word-initial position, with the former appearing in newer layers of the vocabulary and the latter

[19] Tongue root advancement helps create a greater drop in transglottal pressure because it enlarges the supralaryngeal cavity.

[20] Although their presence may be diminishing, as Khachaturian (1992) writes that the breathy-voiced sounds, "though preserved in many cases," may be "undergoing considerable levelling" and are "not consistently present in the pronunciation of many dialect speakers" (127-128).

appearing in older layers (Katvalyan 2016: 98-99; Khachaturian 1992: 124). Since [a] appears after both plain and breathy-voiced consonants (Katvalyan 2016), the lack of breathiness cannot be conditioning the appearance of [ɒ]$_A$. Similarly, [a] appears after [χ] (Katvalyan 2016), which is not aspirated.

Nevertheless, these discrepancies can be explained in a way that is consistent with the theory that the [a]/[ɒ]$_A$ allophony comes from phonologized enhancement of coarticulation. Since word-initial plain voiced stops and affricates are said to have been introduced to the dialect along with newer vocabulary, we can assume that the plain voiced sounds may not have yet been introduced into the dialect at the time when the phonologization took place. Since many word-initial instances of [χ] in the dialect correspond to Classical Armenian [h] (Katvalyan 2016: 102), we can similarly assume that the [a] following word-initial [χ] at one time followed [h] (an environment where [a] currently appears) in many words.[21] This could explain why words that now begin with [χ] were phonologized as environments where [a] appears, if we assume that [a] was historically connected with aspiration/breathiness (and that [h] functioned as an aspirated/breathy sound in the dialect). These indications that the sound system of Gavaṛ's dialect might previously have allowed for the same type of contextual conditioning found by Seyfarth and Garellek (2018) support the idea that the [a]/[ɒ]$_A$ allophony has roots in coarticulation. The fact that the allophony persists today among some speakers despite the possible historical changes to the dialect's sound system supports the idea that the allophony represents a coarticulatory process that was enhanced and phonologized to reflect the phonetic contexts of an earlier period, rather than simply representing synchronic coarticulation of the sort observed by Seyfarth and Garellek (2018).[22]

The above theory does not necessarily account for why nasals behave differently from all other voiced sounds in terms of the contemporary allophonic environments, or for the behavior

[21] Katvalyan (2016: 130) argues, in contrast to previous authors, that these instances of word-initial [χ] do not descend from [h], but rather were present in the relevant dialects already in the pre-Classical period (before 5th century A.D.). If there was a change from [h] > [χ] in the dialect, it indeed must have taken place in a rather early period, preceding the change of word-initial [j] > [h], which took place before the middle Armenian period (before 11th century AD) (see Katvalyan 2016: 127-130). However, given the lack of documentation of dialects in these early periods, it is necessary to consider the explanation for the origin of the [a]/[ɒ]$_A$ allophony which would involve a change from [h] > [χ]. Katvalyan's (2016: 108-109) own explanation for this allophony is that the disparate environments where [ɒ]$_A$ appears are united by "tenseness."

[22] Additional evidence for the enhancement/phonologization of the given process in Gavaṛ is the diphthongization of [ɒ]$_A$ to [oa], which is stronger in the villages surrounding the city (Katvalyan 2016: 67-68). This diphthongization further differentiates this allophonic process from the simple front vs. back variation found in Yerevan.

of the voiced vs. voiceless fricatives (which are followed by [a] and [ɒ]$_A$ respectively). In the case of the nasals, it is possible that the appearance of [ɒ]$_A$ has its roots in a different coarticulatory process, as authors such as Carignan (2018) have found that nasalization can cause lowering of both the first and second formant frequencies (which equates to backing and raising of the vowel). The simple fact that the nasals are not aspirated/breathy could also explain their grouping with the [ɒ]$_A$ environments. This could also explain why [ɒ]$_A$ appears after the voiceless fricatives, but it is unclear if the voiced fricatives (in Gavaṛ or in Armenian in general) have some sort of breathiness that would cause them to group with the [a] environments. Unfortunately, Seyfarth and Garellek's (2018) data does not include word-initial nasals or fricatives, so it is not known if Yerevan speakers show a relatively backed [a] after nasals or a fronted [a] after voiced fricatives, or if their voiced fricatives are associated with any sort of breathy voicing.

As noted above, it is possible that speakers in Gavaṛ who maintain a small difference between [a] and [ɒ]$_A$ are simply the most advanced representatives of a sound change (dialect levelling) in progress. However, it is also possible that they are exhibiting the same type of synchronic coarticulatory effects observed by Seyfarth and Garellek (2018) in Yerevan, and thus that a complete collapse of the difference between [a] and [ɒ]$_A$ would not be expected to take place even if there is a current tendency to shift toward a more merged state. One way to disentangle these two explanations would be to observe whether such speakers produce relatively larger differences between [a] and [ɒ]$_A$ in those contexts that show synchronic coarticulation in Yerevan. However, because Seyfarth and Garellek's (2018) data includes only words that begin with stops, it is not possible to use existing data to compare Gavaṛ and Yerevan speakers across all relevant contexts.

Examining Gavaṛ speakers' data for words starting with [χ] vs. words starting with other consonants or with no initial consonant might also be informative, since we might expect for speakers whose vowel system is characterized by synchronic coarticulation to group initial [χ] (which is not voiced or breathy) with those contexts in which fronting of the vowel is not found. The 2025 word list did include one minimal pair (with four repetitions of each word) that could shed light on this question – *χat͡ʃʰ* ('cross') vs. *at͡ʃʰ* ('right' – dialect pronunciation [ɒt͡ʃʰ]). Visual examination of participants' vowel plots for these words showed that most participants' productions actually did conform with the expectations for the dialect vowel system, in which [χ]

is followed by [a] (i.e., *χat͡ʃʰ* had a higher F2 and also sometimes a lower F1 than *at͡ʃʰ* for most speakers). There were only three participants for whom no similarity to this pattern was identified. It is nevertheless difficult to draw firm conclusions based on a single minimal pair, and it should also be acknowledged that the apparent persistence of the dialectal allophony pattern with respect to the given consonantal context does not necessarily mean that Yerevan's coarticulatory patterns have no relationship with or influence on the speech of those Gavaṛ residents who maintain relatively small differences between the [a] and [ɒ]$_{A}$ allophones. (Relatedly, the apparent persistence of the dialect pattern after [χ] in production does not allow us to rule out influence from Yerevan's coarticulatory patterns on how small differences between these vowels are *perceived* in Gavaṛ.) The lack of data related to this consonantal context in Seyfarth and Garellek's (2018) study also makes it impossible to know what type of vowel is actually found after [χ] among Yerevan speakers. Given these limitations, it is necessary to acknowledge the possibility that the allophonic variation between [a] and [ɒ]$_{A}$ in Gavaṛ is related to one or more coarticulatory processes that are still active in supralocal varieties and could thus also be active for those Gavaṛ residents who do not maintain the dialect's enhanced version of the allophony. The potential influence of the phonological patterns of Yerevan's variety (on both production and perception) thus complicates the endeavor of defining what it would mean for these two vowels to be distinct vs. merged in Gavaṛ. These complexities provide motivation for considering local perception and local social meaning in the study of vowel distinctions and mergers.

Among the various aspects of the Gavaṛ dialect's vowel system, the comparison of [a]/[ɒ]$_{A}$ is also particularly suitable for further investigation because of its apparent salience. Katvalyan (2016: 80) describes [ɒ]$_{A}$ as the dialect vowel system's most salient aspect, and Mkrtch^{h}yan (2016: 203) describes this vowel as what differentiates the dialect from others. This agrees with 2022 participants' metalinguistic commentary, in which back rounded vowels were commented on much more than other dialect vowels. Most of this participant commentary seemed to specifically be about [ɒ]$_{A}$ (rather than [ɒ]$_{O}$), though it is difficult to say this with certainty due to the imprecise nature of much of the metalinguistic commentary. In any case, the salience of [ɒ]$_{A}$ could play an important role in a perception experiment that seeks to consider the potential influence of social information on perception, as Juskan (2016) found social priming

effects in perception experiments only for linguistic variables that were salient in the speech community.

In summary, the 2022 analysis revealed not only correlations between gender and maintenance of some dialect vowels, but also a consistent pattern of differences of varying sizes between [a] and [ɒ]$_A$ that provides evidence for coarticulatory origins of this salient aspect of the dialect vowel system. Together with the fact that much of the social variation in the production data did not seem to have any meaningful relation to the merger/distinction thresholds calculated with the method of Stanley and Sneller (2023), this pattern complicates the already difficult issue of defining what it means for two vowels to be distinct vs. merged. In combination with the evidence for the [a]/[ɒ]$_A$ allophony's connection with synchronic coarticulatory phenomena in Yerevan Armenian, the presence of differences with similar directionality but vastly different magnitudes suggests that a distinction between two vowels can simultaneously be a phonologized, socially salient allophonic phenomenon and a low-level coarticulatory effect within the same speech community (and it is difficult to ascertain where one ends and the other begins).

The 2022 findings thus raise numerous questions: What does it mean for two vowels to be distinct vs. merged in a case where their distinction's potential coarticulatory roots seem to dictate that there might always be some measurable differences in production, even if these are not perceptible (in general, or for some individuals / under some circumstances)? How does this case complicate previous findings about near-mergers between vowels? Are there instances in which vowels should be considered merged / functionally the same (rather than nearly merged) despite the existence of consistent measurable acoustic differences? Where is the line between low-level coarticulatory effects and phonologized, socially salient allophony (especially when these might coexist along a continuum within the same community)? How should we consider acoustic measurements and local perception of differences in defining whether two vowels are distinct vs. merged, and how should we interpret cases where these two metrics conflict? Given that the allophonic phenomenon under consideration is socially salient, how might social information affect its perception? In general, how should we incorporate the widespread finding that social information can influence perception into our understanding of what it means for vowels to be perceived as distinct or merged in a given community? These issues will be explored through the specific research questions outlined above in Section 3.2.

3.4 Motivations from previous literature

The 2025 study builds upon previous research related to vowel mergers and near-mergers, relationships between speech production and perception, perception of coarticulation and of allophonic variation, and the role of social information in speech perception. This dissertation contributes to the previous literature by addressing questions that lie at the intersection of these topics and bring together issues raised by various previous studies. The following sections will review literature related to these topics, highlighting the ways in which this literature informs the research questions of the present study. Overall, engaging with issues raised by this literature furthers the study's goal of investigating the influence and interaction of social and phonological factors on speech perception and production, as well as the goal of considering how social, phonological, and individual factors should be considered in defining vowel distinction vs. merger.

3.4.1 Vowel mergers, near-mergers, and production-perception links

3.4.1.1 Defining merger and near-merger, conceptually and quantitatively

In the study of languages' vowel systems, the term "merger" is typically used to refer to situations in which a phonemic or contrastive distinction is lost (see Gordon 2013; Labov 1994: 324; Maguire et al. 2013), either in general or in particular phonetic contexts. In contrast, the vowel pairs under consideration in this study are either fully or partially allophonic – [a] and [ɒ]$_{A}$ are in a completely allophonic relationship conditioned by phonological environment, while the relationship between [ɔ] and [ɒ]$_{O}$ is contrastive but conditioned by environment and stress in some subsets of vocabulary, likely a remnant of an allophonic relationship that existed at some previous time. Although the literature on variation in the production and perception of allophonic differences is comparatively small (see Section 3.4.2), the literature on phonemic vowel mergers provides many relevant insights, particularly regarding relationships between production and perception of vowel differences.

A very widespread finding from previous work on vowel mergers is the mismatch of production and perception. Speakers have been found to produce phonetic differences between vowels that they and/or other members of their community do not perceive (e.g. Arnold 2015; Austen 2020; Benus & Gafos 2007; Bowie 2000; Costa & Mattingly 1981; Di Paolo 1992; Di Paolo & Faber 1990; Freeman 2023; Herold 1990; Janson & Schulman 1983; Kontra 1993;

Labov 1994; Labov et al. 1972; Labov et al. 1991; Labov et al. 2006: 62; Strandberg et al. 2022). Such instances often involve relatively small acoustic differences and are often referred to as "near-mergers" or "apparent mergers." Similar phenomena – in which small phonetic differences are found despite the assumption that sounds are merged – have also been observed for other types of sounds. Di Paolo and Faber (1990: 157) and Benus and Gafos (2007: 288) in discussing cases of apparent or near vowel merger, compare the situation to cases of "incomplete neutralization" of voiced and voiceless word-final obstruents observed in various languages (for discussion of incomplete neutralization related to voicing and other consonant phenomena, see Beaton [2016]; Kharlamov [2014]; Kleber et al. [2010]; Nicenboim et al. [2018]; Oh [2022]; Port & Crawford [1989]; Roettger et al. [2014]; Warner et al. [2004]).

Previous literature shows a great deal of disagreement regarding how near-merger (and even merger) should be defined conceptually and quantitatively. In reflecting on the impreciseness of the definition of merger, Maguire (2007: 227) points out that phenomena referred to as mergers are usually "gradient, and may correlate, to varying degrees, with a variety of social factors." Maguire (2007: 62) identifies a continuum from "complete, language-wide mergers" to "variable mergers and near mergers" and argues (2007: 350-351) that merger should be defined at the individual rather than community level due to the existence of vast phonological variation among speakers. Maguire (2007: 44) also points out that the gradual nature of merger by approximation makes it difficult to identify the exact point at which the merger has been completed, and that such a process can result in near-merger if it halts "at the last moment" (2007: 325) and results in "subtly (but significantly) different phonetic distributions" (2007: 319). This discussion raises questions about how to define a subtle but significant acoustic difference between vowels.

While the perception-production mismatch described above is surely crucial for identifying a given phenomenon as a near-merger, various authors have also attempted to apply purely production-based acoustic criteria in defining near-merger and merger. Recasens and Espinosa (2009: 241) use "just noticeable difference" data reported by Boersma (1998: 104) to come up with a near-merger threshold in Hz for the mid vowels under examination in their study. Kotsoni (2019: 64-65) comes up with Hz value cutoffs for merger, near-merger, and contrast, citing Labov (1994) and Recasens and Espinosa (2009). Such methods contain an element of arbitrariness, in that they pair acoustic thresholds with phonological states (e.g., merged, near-

merged, contrastive) without sufficient external justification for why the given acoustic thresholds should be viewed as equating to the given phonological states. When these methods rely on "just noticeable differences" in particular, they conflate acoustic discriminability under ideal listening conditions with more phonologically mediated perception of vowels. For example, Recasens and Espinosa (2009) define near-merger as a difference that is smaller than the "just noticeable difference" reported by Boersma (1998: 104). However, such a method for defining near-merger is questionable, as it does not seem that previously described cases of near-mergers always involve acoustic differences that are small enough as to be generally indiscriminable. Indeed, a defining characteristic of near-mergers is the fact that they often involve differences that are large enough to be perceptible (i.e., discriminable and/or able to be used for semantic categorization) to outsiders, but are not perceptible to members of the speech community in question (e.g., in the case of the *ferry/furry* merger / near-merger in Philadelphia studied by Labov et al. [1991]). Thus, it would seem that attempting to use arbitrary (e.g., derived from studies whose research questions are very different from those explored in studies of vowel mergers) acoustic thresholds to define near-merger (or other phonological concepts) is inappropriate.

Previous studies of mergers have also used various statistical significance tests as quantitative definitions of merger in production. These include p-values from the MANOVA output (for Pillai scores) (e.g., Bowie 2021; Sloos 2013), significance testing of F1 and F2 differences using ANOVA (e.g., Arnold 2015), and t-tests (e.g., Chevalier 2016). While Stanley and Sneller's (2023) method improves upon these by controlling for varying sample size, the problem of the arbitrariness of statistical significance cutoffs remains. Strelluf (2014: 91), in citing Herold (1990) and Johnson (2010), notes that such authors have often found statistically significant differences in the English low back vowels among speakers whom they deemed merged, which would suggest that statistical methods are sometimes too sensitive to small (perhaps not meaningful) acoustic differences. On the other hand, one could ask whether the given authors simply failed to identify cases of near-merger, in which the speakers maintain a difference in production but not perception. Evanini's (2009: 141-142) discussion of the English low back merger sheds some light on this question and further complicates the definition of near-merger. In discussing results of an analysis that included perceptual results (from minimal pair tests) for some but not all speakers, Evanini points out that the Euclidean distance between the

vowel classes involved in the low back merger "rarely approaches 0, even for a completely merged speaker, due to the different historically-derived allophonic distributions of the two classes" (2007: 141). Evanini (2007: 141-142) clarifies that both merged and unmerged speakers have similar directionality of differences in F1 and F2 between the two vowel classes. While one might again argue that Evanini is simply describing cases of near-merger, Evanini's attribution of differences in phonetic distribution to "different historically-derived allophonic distributions" calls into question whether a phenomenon that initially appears to be a near-merger should always be classified as such. In other words, if we assume that factors such as differing consonantal environments associated unevenly with different vowel classes can inhibit a total phonetic/acoustic merger of those classes, it becomes less clear what should be considered a true near-merger (in which speakers are mysteriously maintaining a difference that they cannot perceive) vs. what should be considered a natural remnant of previous allophony and/or a natural result of coarticulatory effects of unevenly distributed neighboring consonants. Given the nature of the vowel pairs investigated in this dissertation, this question is very relevant in determining the degree to which the given vowel phenomena found in Gavaṛ can be likened to near-mergers. In any case, while a mismatch between perception and production is clearly key to the definition of near-merger, it is less clear what type and size of difference in production is needed in order for such a mismatch to be said to exist.

3.4.1.2 Complications related to the role of perception-production relationships in vowel mergers and near-mergers

When near-mergers are found in situations that are thought to be mergers in progress, it is often said that production of the vowels is lagging behind their perception. However, the opposite type of mismatch, in which listeners can perceive differences between vowels that they produce as merged (i.e., perception lagging behind production), has also been found (Austen 2020; Baranowski 2013; Freeman 2023; Hay et al. 2006; Labov et al. 2006: 62; Thomas and Hay 2005; Zhang et al. 2022). Though the existence of this phenomenon has sometimes been attributed to orthographical differences between words (e.g., Labov 1994: 357; Labov et al. 2006: 29), other authors (e.g., Babel et al. 2013; Freeman 2023; Hay et al. 2006; Thomas & Hay 2005; Zhang et al. 2022) suggest that interaction with speakers who produce unmerged variants is responsible for the maintenance of distinctions in perception. Freeman (2023: 2) and

Baranowski (2013: 281-283) also connect this phenomenon with stigma or negative social evaluation associated with given vowel mergers.

The findings reviewed above may seem contradictory, as some suggest that perception leads production during mergers in progress and other sound changes, while others suggest that production can be ahead of perception (some studies also find alignment of production and perception during sound change; see Harrington et al. [2019] for an overview). However, a more careful consideration of the previous evidence reveals that different types of mismatches tend to occur systematically in different types of situations. Importantly, some of the studies listed above (e.g., Austen 2020; Freeman 2023) find evidence for both types of mismatches. In the case of Freeman's (2023: 16) study of a vowel merger that could be reversing, the different types of mismatch are found within the same region, with younger participants producing differences that they have difficulty perceiving, and older participants perceiving differences that they do not necessarily produce. Di Paolo (1988: 91), in examining several vowel mergers and shifts in progress, finds an opposite (but similar) phenomenon – perception leads production among older speakers, but for younger speakers "pronunciation either equals or passes up categorization." A similar result was found by Pinget et al. (2020: 660), who studied two changes involving consonant devoicing in Dutch and found that perception leads at the beginning of a sound change but lags behind production as the change nears completion, such that "individuals still perceive a contrast they no longer produce." Similarly to some of the authors mentioned above, Pinget et al. (2020: 678) also note that speakers who no longer produce a contrast may need to retain perceptual knowledge of it in order to understand speakers who still produce it.

This type of explanation for perception/production mismatches can also be invoked in relation to vowel phenomena other than mergers. For example, Janson (1983: 31) finds perception to lag behind production in a vowel shift across generations and points out that the younger generations must retain the ability to correctly classify older generations' pronunciations. Kettig and Winter (2017), in finding more socially structured variation in production than in perception of the Canadian Vowel Shift, similarly suggest that listeners must perceptually accommodate to both innovative and conservative variants, regardless of their own productions. Additionally, D'Onofrio's (2021) findings related to production and perception of younger speakers who are reversing the Northern Cities Vowel Shift are similar to Freeman's

(2023) finding that younger participants' perception of the LEG-VAGUE merger aligns more closely with older speakers' productions than with their own.

As all of these findings demonstrate, a variety of factors can influence whether a given speaker produces vowel distinctions they do not perceive or perceives vowel distinctions they do not produce, with these mismatches being manifestations of broader tendencies that can also occur in other types of sound changes as they progress across generations. As Section 3.4.3 below will explore, the relationship between production and perception of vowel mergers is further complicated by the fact that listeners' perception of vocalic contrasts also cannot be assumed to be fixed synchronically, as perceived social information about the talker has been found to affect listeners' perceptual judgements of vowel mergers (e.g., Drager 2005; Hay et al. 2006; Koops et al. 2008).

3.4.1.3 Mixed and contradictory results of more general investigations of perception-production links

The various relationships between perception and production uncovered in studies of vowel mergers and other vowel shifts must also be considered against the broader backdrop of research into individual and group level links between speech perception and production, which has shown very mixed results. Some studies have found relatively straightforward relationships between corresponding aspects of individuals' production and perception. For example, Fox (1982) found correlations between aspects of individuals' acoustic vowel spaces and their perceptions of vowels, while Bell-Berti et al. (1979) found two different articulatory strategies for the production of front vowels among participants, which corresponded to differences in the perception of vowel boundaries. Newman (2003) found correlations between production and perception of consonants, for example identifying that individuals who produced longer VOT in stop consonants also had longer VOT in their perceptual prototypes. Evans and Iverson (2007) similarly found that participants' vowel productions were similar to their perceptual prototypes, while Perkell et al. (2004) found that participants who produced more acoustically distinct vowels were also better at discriminating vowels in perception. Though studies of vowel mergers in progress have often found complex relationships between production and perception, as described above, more straightforward relationships are also sometimes found. For example, Hay et al. (2006) found that a participant's degree of merger in production predicted their performance in a lexical identification task, while Loakes et al. (2014) found an association

between participants' production status (merged vs. unmerged) and their performance in an identification task. Kendall and Fridland (2017) found that individual participants' degree of participation in the cot-caught merger predicted their perception of another nearby boundary in the vowel space – namely, the boundary between /æ/ and /ɑ/. Some studies have also found relatively straightforward relationships between production and perception at the level of broad demographic groups. For example, Thomas (2000) found that dialectal variation in the production of sub-phonemic vowel differences was reflected in variation in the use of those differences as a perceptual cue. Some studies have also found that broad demographic categories and characteristics of individual speech production are both relevant for speech perception. For example, Fridland and Kendall (2012) and Kendall and Fridland (2012) both found that both regional affiliation and aspects of individuals' vowel productions were related to their perception of vowels. Thus, many studies have suggested that aspects of speech production and perception are correlated at both the individual and group levels.

Various other studies have complicated or contradicted these results. Despite identifying links between production and perception of several phenomena (e.g., coarticulatory nasalization, reliance on VOT vs. f0 cues), Beddor (2015) also notes the existence of individuals whose production and perception are not neatly correlated for a given phonetic property, especially emphasizing the flexibility of perception and the tendency for certain individuals to consistently perceive cues that they do not produce. In general, some studies have found that relationships between production and perception can vary greatly across individuals. For example, Frieda et al. (2000) found that participants' hyperarticulated vowels (rather than vowels from citation speech) aligned more closely with their perception of those vowels (see Johnson et al. [1993] for discussion of the "hyperspace effect"), but this was only true for participants who had clear perceptual prototypes. Some studies have sought and failed to find any correlations between individuals' production and perception of various types of sounds, including stop consonants (Bailey & Haggard 1973), vowels (Paliwal et al. 1983), and glides (Ainsworth & Paliwal 1984). Another study by Mitterer and Ernestus (2008) suggests that a link between speech production and perception exists, but argues that this link is abstract and phonological rather than related to phonetic detail. Similarly, Kraljic et al. (2008) find that speakers can learn to perceptually accommodate to the phonetic details of an unfamiliar idiolect without changing their own productions. These findings align with a point made by Thomas (2000: 19), who writes that

"listeners can acquire the ability to recognize phonetic cues that they do not produce when they are in contact with speakers who do produce those cues," and of Sumner and Samuel (2009), who suggest that an individual can be a "fluent listener" of a dialect that they do not produce. Though comparing production and perception in studies of vowel mergers has often revealed various types of correlations, mismatches, and social structure, some studies have found less clear results in this regard. For example, Hay et al. (2013) found no correlations between distinctiveness of individuals' productions of a targeted distinction and their accuracy in lexical identification tasks. Warburton (2020: 189-203), in examining a situation involving vowel overlap that she ultimately hesitates to call a merger in progress, found that some participants' productions and perceptions were approximately aligned, some had a more advanced merger in production, and some had a more advanced merger in perception. However, unlike some of the authors mentioned above, Warburton did not find that these various production-perception relationships had a clear social pattern.

As this literature review demonstrates, the study of individual- and group-level relationships between speech production and perception has shown extremely varied results both in the case of vowel mergers and in more general investigations of perception-production links. In any case, it is clear that correlations between aspects of perception and production, as well as mismatches in which perception either leads or lags behind production, are all possible scenarios. As some of the studies reviewed above suggest, factors such as the completion status of a sound change or an individual's age may help to explain when and why certain perception-production relationships are found, although there is clearly still a great deal to be learned about the systematicities that may govern perception-production relationships.

3.4.1.4 The importance of considering different perceptual behaviors in the study of vowel mergers and near-mergers

Some of the earlier studies of vowel mergers and near-mergers draw attention to the crucial idea that the production of vowels can relate to different perceptual behaviors in different ways. Labov et al. (1972: 256) emphasize the issue of the types of tasks used to explore near-mergers, pointing out that the labeling tasks used in their studies probe listeners' abilities to use acoustic differences for semantic categorization, rather than the more basic ability to discriminate a difference in sound. In order to address this issue, the authors call for discrimination tasks with synthetic speech stimuli to be used in addition to labeling tasks. Herold

(1990: 173) also draws attention to this issue and discusses results that show that "the ability to perceive a difference in sound does not imply the ability to use that difference to identify lexical items," and that "failure to identify lexical items correctly does not imply an inability to hear the differences between them." In studying the near-merger of the vowels of "ferry" and "furry" in Philadelphia, Labov et al. (1991: 71-72) found that some of the small acoustic differences maintained by Philadelphians between these two word classes were very difficult to categorize not only for other Philadelphians, but even for non-Philadelphians, for whom these two word classes were not undergoing merger. In discussing this somewhat surprising result, which shows that speakers can consistently produce differences that are small enough to cause categorization difficulties even for listeners with well-separated phonological categories, Labov (1994: 418) suggests that Philadelphians' psychoacoustic discrimination abilities may explain how they are able to learn and consistently produce a very small acoustic distinction that they do not use consciously for semantic categorization. In other words, as Eckert and Labov (2017: 486) emphasize, judgements of pronunciations as the "same," often indicate "a suspension of the labeling function" rather than an inability to discriminate an acoustic difference.

Thus, in comparing production and perception in vowel mergers and near-mergers, it is important to consider not only semantic categorization (commonly explored in studies of vowel mergers through methods such as minimal-pair tests and commutation tests [see Labov [1994]] and other vowel identification tasks), but also other perceptual abilities such as discrimination. Accordingly, the misalignment of production and semantic categorization of merging vowels (as is found in near-merger) does not necessarily imply that production will be misaligned with other perceptual behaviors like discrimination. Indeed, several early studies of near-mergers also emphasize that speakers generally acquire and produce phonetic detail that they are unaware of and/or do not use for semantic categorization, such that the maintenance in production of an acoustic distinction that is not used for categorization should not necessarily be surprising. For example, Labov et al. (1972: 254) emphasizes that speakers subconsciously acquire a great deal of phonetic detail that characterizes their own dialect or vernacular, while Herold (1990: 214-215) notes that speakers acquire language-specific allophonic patterns that are not used for semantic distinction. In discussing Labov's studies of near-mergers, Nunberg (1980: 229) proposes how speakers may be able to produce differences that they are not aware of being able to hear, emphasizing that "two points might be discriminable for a child learning to produce the

phones of his language, and at the same time not discriminable to an adult who is casually asked to identify two tokens." Janson and Schulman (1983: 334), in discussing a group of Swedish speakers who maintain a vowel distinction that they no longer use to categorize words, also compare this situation to subconsciously acquired allophonic variation, proposing that such variation conveys information about the speaker's dialect or social characteristics rather than semantic information. In discussing near-mergers as a general phenomenon, Gordon (2013: 208, 218) contrasts them with phonetically conditioned allophonic differences, describing the latter as a case where speakers' consistent production of differences that they do not "perceive" is normal and widespread. Overall, these studies highlight the importance of considering different types of perceptual behaviors in studies of vowel mergers and near-mergers, emphasizing that speakers' productions of different vowel classes should be compared not only with their conscious categorization of them, but also with their discrimination of the relevant acoustic differences, which may be below the level of awareness and/or not have a semantic differentiating function.

3.4.1.5 Contributions of the perception experiment to the study of perception-production links

The present study contributes to the literature reviewed in the previous sections not only by providing additional data from an understudied language variety, but also by engaging with the issue of how production might relate to different perceptual behaviors, which has not always been thoroughly investigated in studies of vowel mergers and near-mergers (despite the abovementioned authors' acknowledgement of its importance). In particular, the present study combines consideration of this issue with consideration of perception of allophonic differences. Although several of the studies cited above liken allophonic differences to those that listeners must physically hear but are not conscious of (i.e., are unable to successfully use for semantic categorization), the present study considers a situation in which allophonic differences are likely to carry social salience, which adds an additional layer of complexity to the issue of considering differences between different perceptual behaviors. The present study thus assumes that socially mediated perception of allophonic differences – like the use of acoustic differences for conscious semantic categorization – will not necessarily align with simple discrimination of acoustic differences, separating discrimination and the use of allophonic and other differences for dialect categorization into separate tasks. Given that the study involves a potentially socially salient allophonic difference, which is also potentially rooted in coarticulatory processes, the following sections review relevant literature related to perception of allophones, perception of

coarticulation, and the influence of social information on perception, before situating the present study in relation to this previous literature.

3.4.2 Perception and production of coarticulation and allophonic variation

Like the literature about perception-production links in vowel mergers and in general, the literature related to perception of non-contrastive phonetic detail, such as coarticulatory effects and allophonic variation, also presents conflicting theories and results. As the vowel systems examined in the present study potentially simultaneously involve synchronic coarticulatory effects and allophonic distinctions that have emerged diachronically, the study can contribute to the understanding of whether and how listeners use different kinds of non-contrastive phonetic detail in perception.

A multitude of studies have suggested that listeners perceptually compensate for coarticulation, effectively factoring out the phonetic influence of a certain sound on adjacent sounds and attributing this influence to its source (see Fowler 2006; Mann 1980; Mann & Repp 1980; Mann & Repp 1981; Mitterer 2006; Repp & Mann 1981; Viswanathan et al. 2010). For example, in hearing a vowel that is nasalized due to the coarticulatory effect of a following nasal consonant, listeners will attribute the nasalization to this consonant and will not perceive the vowel as nasalized. Other studies have complicated this assessment of the perception of coarticulation, finding that perceptual compensation is only partial (Beddor & Krakow 1999) and that listeners do not merely factor out coarticulatory influence, but also can use it as a clue for identifying adjacent segments (see Beddor et al. 2013; Fowler 2006; Repp & Mann 1981). Perceptual compensation for coarticulation has been found to vary across languages (Beddor et al. 2002; Beddor & Krakow 1999; Kang et al. 2016), mirroring language-specific coarticulatory patterns (Beddor et al. 2002).

Some studies have also found individual variation in both coarticulation in speech production and sensitivity to coarticulatory effects in speech perception (e.g., Beddor et al. 2018; Grosvald & Corina 2012; Zellou 2017), though these studies' findings differ regarding whether coarticulation in production and use of coarticulatory effects in perception are correlated (Beddor et al. 2018; Coetzee et al. 2022; see also Beddor 2015), correlated in some tasks but not others (Zellou 2017), or not correlated (Grosvald & Corina 2012). Variation in the perception of coarticulatory effects has been proposed to be linked to sound change. For example, Ohala (1981) proposes that sound changes arise when some listeners fail to perceptually compensate

for coarticulation, instead perceiving coarticulatory effects as the speaker's intended phonetic target, and various studies (e.g., Kleber et al. 2012; Kuang & Cui 2018) provide support for this theory. Together with the varied findings regarding individual correlations between production and perception of coarticulation, these findings suggest that production-perception mismatches related to coarticulatory effects are possible, just as in the other domains reviewed above, though more research is needed to understand why various types of correlations and mismatches are found in different studies. It seems possible that social factors could play a role in mediating perception of coarticulation, given the general body of literature about the effects of social factors on speech perception (see Section 3.4.3). However, the potential role of social factors has not been extensively explored. Coetzee et al. (2022) found that information about the ethnic identity of the talker did not affect listeners' reliance on coarticulatory cues in perception (despite these coarticulatory cues differing in the productions of different ethnic groups), but further research is needed to investigate the potential effects of a wider variety of social factors.

Studies that directly compare perception of allophonic variation with perception of phonemic differences have generally found that discrimination of allophonic differences can differ from discrimination of phonemic differences, either being less or more easily discriminable according to different studies. The most general manifestation of this tendency is found in the phenomenon known as "categorical perception" (see Liberman et al. [1957] for one of the earliest studies of this phenomenon and McMurray [2022] for a more complete overview of research on categorical perception), in which listeners are found to more accurately discriminate differences that are near phonemic boundaries than those that are not. Accordingly, response patterns found in relation to the same set of stimuli can vary in relation to listeners' language backgrounds (specifically, in relation to which sounds are contrastive in their language(s)). For example, Amengual and Chamorro (2015), in examining Galician-dominant and Spanish-dominant listeners' discrimination of vowels that varied along continua and are contrastive in Galician but not in Spanish, found differences in response pattern based on the listeners' dominant language. In particular, Galician-dominant listeners showed peaks in the accuracy of their discrimination at certain points along the continua, while Spanish-dominant listeners had overall lower accuracy and flatter response patterns.

Lower accuracy of discrimination of allophonic differences has been found in several other studies. For example, Peperkamp et al. (2003) found that French speakers discriminated

allophonic consonant differences with more difficulty than phonemic differences, particularly when embedded in their usual phonetic context, while Boomershine et al. (2008) found that English and Spanish speakers were more successful at discriminating consonantal distinctions that were contrastive in their language, when compared with allophonic distinctions. Nevertheless, Rose et al. (2023) found that an allophonic vowel difference was better discriminated by Akan speakers than a phonemic difference in which the two vowels were more acoustically similar, suggesting that the finding of more difficult discrimination may not generalize to all cases of allophony. Importantly, Peperkamp et al. (2003), Boomershine et al. (2008), and Rose et al. (2023) all used stimuli that were nonsense words, VCV sequences, or isolated syllables, and sometimes also involved telling participants that the stimuli were from a foreign language. It is thus not entirely clear if the findings are comparable to perception of allophonic variation in more natural speech contexts, such as those used in the present study. Thomas (2000) used more natural stimuli, and as noted above, this study found that dialectal variation in the production of sub-phonemic vowel differences aligned with variation in the use of those differences as a cue in perception. This suggests that the perception of allophonic variation may somehow be correlated with its production in some situations.

Mismatches in the production and perception of allophony have also been found, with much relevant evidence coming from studies reviewed above, which examine phonetically conditioned vowel mergers (i.e., neutralization of vowel contrasts that are restricted to certain phonetic contexts). Such studies include Austen (2020), which examines the PIN-PEN merger in US English, and Freeman (2023), which examines two manifestations of pre-velar vowel raising – the LEG-VAGUE merger and BAG-raising – in Pacific Northwest English. Both studies found different types of production-perception mismatches. For example, as noted above, Austen (2020) found that perception was ahead of production for some participants and lagged behind production for others. Freeman (2023) found that the LEG-VAGUE merger was more complete in perception than in production (near-merger), while BAG-raising was more prominent in production than in perception, such that "most participants did not perceive BAG-raising, regardless of their own production" (Freeman 2023: 15). Freeman notes that the LEG-VAGUE merger is a more advanced sound change, while BAG-raising is more variable, socially salient and possibly stigmatized, and this difference is reflected in the results of the study. Specifically, the LEG-VAGUE merger (less socially salient/stigmatized) is perceived as more advanced than

it actually is, while BAG-raising (more socially salient/stigmatized) is perceived as less advanced than it actually is. Importantly, the pre-velar raising phenomena examined by Freeman likely have coarticulatory origins (see Mielke 2017), which highlights the connection between coarticulatory effects and the diachronic development of allophony. Thus, both Austen and Freeman's studies demonstrate the existence of production-perception mismatches in phonetically conditioned vowel phenomena, and Freeman's work in particular highlights the links between such phenomena and other types of non-contrastive phonetic detail (i.e. coarticulatory effects), as well as the potential importance of social meaning in mediating perception and production of such phenomena.

Overall, it seems that listeners do sometimes attend to allophonic variation, despite the previous findings that allophonic differences are often difficult to discriminate. This is not particularly surprising given Labov's (1994: 324) assessment that "Listeners attend to the realization of particular sounds by other speakers, but not to how many distinctions they make" and Eckert and Labov's (2017: 467) argument that social meaning is assigned to "phonetic elements" rather than "to the phonological structures in which those sounds participate." Kristiansen (2003: 111-112) likewise argues that allophonic variants can convey social information such as linguistic stereotypes. These statements suggest that contrastiveness is not particularly important for a phenomenon to be perceived and/or assigned social meaning, despite some authors' (e.g., Hickey 2002: 119; Trudgill 1986: 11) assertions that contrastive and/or phonetically unconditioned phenomena are more likely to be salient. A review of the relevant literature on perception of coarticulatory and allophonic phenomena demonstrates that the perception of both types of phenomena is highly variable, with this variation depending on factors such as the particular phenomenon under examination (e.g., its social salience, whether or not it is currently undergoing a sound change, its phonetic and phonological characteristics), the task type, listeners' language background, and idiosyncratic individual differences. Though some work has found correlations between production and perception of allophony and between individuals' production of coarticulation and perceptual use of coarticulatory information, other studies of coarticulation have found mismatches between these metrics, and studies of allophonic variation (in the form of phonetically conditioned vowel mergers, which form a subset of the mergers examined in the studies reviewed above) have also often found misalignment of perception and production.

Though the influence of factors such as language background on the perception (and sometimes also production) of non-contrastive phonetic detail has been relatively well examined by the studies reviewed above, the influence of a variety of social factors – such as social salience of a given phenomenon or social framing of a task – is less clear, especially in the case of coarticulatory phenomena. Given that coarticulatory and allophonic phenomena are related, such that in many cases (such as those examined by Freeman [2023]) it may be difficult to clearly separate them, it is interesting to consider the extent to which social salience – or, at the very least, some sort of conscious or unconscious noticing of a phenomenon – could play a role in the enhancement of coarticulatory effects and the diachronic development of allophony. For example, does salience make it more likely that individuals will attend to and use non-contrastive phonetic detail in perception, perhaps also exaggerating it in production over time? Could tasks that draw more or less attention to the social meaning or salience of non-contrastive phonetic detail influence the degree to which listeners attend to this detail? The present study considers these issues through the examination of pairs of vowels with different characteristics, and through the use of different types of tasks, which is also motivated by the literature discussed in Section 3.4.1 above ad Section 3.4.3 below.

The vowel systems under consideration in the present study provide a particularly interesting case for the examination of these issues because the [a]/[ɒ]$_{A}$ comparison simultaneously represents an allophonic phenomenon that has emerged diachronically in Gavaṛ's dialect and a synchronic coarticulatory process in the supralocal variety with which it is in contact. Because of this, its production and perception by participants can be expected to vary widely, and different task types may be able to access different modes of perception (i.e., perception of this phenomenon as a synchronic coarticulatory effect vs. perception of it as dialect-specific, socially salient phenomenon). In contrast, the [ɔ]/[ɒ]$_{O}$ vowel pair is less socially salient, not entirely allophonic, and not clearly related to any coarticulatory process. Comparing the results for these two vowel pairs will thus generally shed light on the influence that some phonological and social factors might have on perception and production, and examination of the perception of the [a]/[ɒ]$_{A}$ pair in particular will help to address the relative lack of previous studies related to the influence of social factors on perception of non-contrastive phonetic detail.

3.4.3 The role of social information in speech perception

The use of multiple different perception tasks in the present study has already been motivated above by the need to explore different perceptual behaviors and by the goal of attempting to access different modes of perception of non-contrastive phonetic detail. However, additional support for this methodological decision comes from the literature related to the influence of social information on speech perception. Though this literature generally does not examine allophonic (i.e., phonetically conditioned) vowel phenomena of the type considered in the present study, its findings bolster the assumption that social information is probably also relevant in the perception of such phenomena.

Many past studies have found that various types of social information can affect speech perception. Apparent/perceived social factors that have been found to influence perception include the talker's regional origin (Niedzielski 1999; Schertz et al. 2019), age (Hay et al. 2006; Koops et al. 2008), social class (Hay et al. 2006), and gender (Johnson et al. 1999; Munson 2011; Strand 1999). Thus, in designing tasks for a perception study, it is necessary to consider the potential effects of various types of social prompting. A few studies have found speech perception differences between conditions where listeners are given information about the talker's identity in comparison with a baseline condition (D'Onofrio 2018; McGowan 2015), and a few others have found that the presence vs. absence of social information about the talker affects speech accommodation (M. Babel 2009, 2012). These findings suggest that comparing tasks with and without explicit social information could be an effective way of accessing different possible types of perception of vowels in Gavaṛ, such as the discrimination of small acoustic differences in vowels and the use of vowels for social categorization of the talker. Comparing such tasks could be especially informative in the examination of the [a]/[ɒ]$_{A}$ vowel pair, in connection with the different possible sources (coarticulatory vs. socially salient allophony) of differences between these two vowels.

There have also been some studies that have investigated linguistic perception and/or social evaluation of talkers in tasks with more vs. less explicit social information about the talker, which have found differing outcomes in more implicit vs. more explicit conditions (e.g. Campbell-Kibler 2012; Mack & Munson 2012; Munson et al. 2017; Squires 2013). Related arguments and findings have been discussed by many other authors. For example, A. Babel (2016) proposes different levels of sociolinguistic awareness and control, Preston (2016) argues for the separation of implicit and explicit reactions to a linguistic stimulus, and Campbell-Kibler

(2016) highlights the difference between unconscious and conscious sociolinguistic behavior. McGowan and A. Babel (2019) found that listeners' explicit social evaluations of a talker diverged from the results of their performance in a perception task, while D'Onofrio (2018) found different results based on the task type and time course of the task when probing implicit sociolinguistic awareness, suggesting different types of awareness even within the unconscious level. Squires (2016: 64), in discussing differences between "low-level automatic perception" and "conscious noticing," summarizes that "perceiving linguistic differences might not always lead to noticing, and explicit awareness of differences… may affect what is perceived." Such arguments and findings provide further support for the hypothesis that the presence or absence of explicit social information could affect perception of vowels and vowel boundaries in Gavaṛ, as these different task types could access more vs. less conscious types of perception.

3.4.4 Summary of contributions of the perception experiment

Given the questions raised by the previous literature reviewed above, the contributions of this study can be summarized as follows:

1) Considering that different perceptual behaviors have not always been carefully separated in previous work on perception and production-perception relationships (particularly in the case of work on vowel mergers), the present study will contribute to increased clarity in this issue by investigating perception and possible perception-production relationships under various circumstances. This is accomplished through the use of multiple perception tasks that probe different perceptual abilities, and through examination of the relationships between speech production and the outcomes of these different perception tasks. Though previous discussions of this issue in the literature on vowel mergers have focused on separating acoustic discrimination from successful semantic categorization, the present study instead separates acoustic discrimination from evaluation of dialect association, in accordance with the largely non-contrastive vowel phenomena under consideration. In so doing, the study considers not only potential differences between different perceptual behaviors, but also potential effects of social prompting in the task, aligning with the study's broader goal of considering the influence of social factors on perception of vowels.

2) The project contributes to the literature on perception of non-contrastive phonetic detail (such as coarticulatory effects and allophony), specifically by considering the potential effects of social salience of the phenomenon and of social prompting in the task on the

perception of such phenomena. Previous work on the effects of social factors on the perception of such phenomena is sparse, especially in the case of coarticulatory effects, but considering such factors more thoroughly could potentially shed light on some of the contradictions in the findings of previous studies in this area. By implementing tasks that do or do not draw attention to social factors and examining phenomena that are ambiguous in their status as synchronic coarticulation or diachronically solidified allophony, the present study will seek to identify if perception of non-contrastive phonetic detail and its relationship to production varies according to which characterization of the given phenomena (synchronic coarticulation or diachronically solidified and socially salient allophony) is likely to be activated by a given task. More generally, by comparing pairs of vowels that differ in their phonological status, the study will also shed light on how perception of non-contrastive phonetic processes might differ from perception of contrastive ones. In implementing different types of perception tasks and examining pairs of vowels with differing characteristics, the study will thus shed light on the broader question of how social and phonological factors influence perception and perception-production relationships.

3) In connection with the particular characteristics of the vowel phenomena under consideration, the present study will also contribute to discussions of how to define vowel mergers and near-mergers. The present study differs from the studies of vowel mergers reviewed above in that it involves a phonetically conditioned allophonic vowel phenomenon that has no connection with context-dependent phonemic neutralization. This phenomenon is also socially salient and potentially related to a synchronically active coarticulatory process. This type of allophony, which is not represented in the previous literature on vowel mergers and near-mergers, creates several complications for defining concepts such as "merger," "distinction" and "perception-production mismatch." Thus, the present study will demonstrate the importance of defining such concepts more carefully and in ways that are context-dependent.

As most previous studies on near-mergers by definition focus on phonemic vowel mergers, it is difficult to say if small, non-perceived differences between the vowels involved in those cases are functionally comparable, for example, to the small differences in [a]/[ɒ]$_{A}$ produced by speakers like Speaker 4 (see **Figure 3.5**). Since small differences like the one produced by Speaker 4 could be due to the synchronic coarticulatory process discussed above, it is not clear if we should expect anyone (regardless of their degree of adherence to the dialect

vowel system) to differentiate them (in general or in certain types of tasks), since perceptual compensation for coarticulation could be occurring to some extent. Thus, although the previous literature tells us that production/perception mismatches frequently occur, such that people produce phonetic differences that may be too small for they or others to discriminate, in cases such as [a]/[ɒ]$_A$ in Gavaṛ, the size of the difference that might be necessary in order to be perceived in various types of tasks might be different than for phonemic mergers.

These considerations complicate our understanding of what it means for there to be a mismatch between production and perception of a pair of vowels, adding to the previously discussed implications of considering different perceptual behaviors for this issue. For example, if speakers consistently produce a small, measurable difference between [a]/[ɒ]$_A$ because of coarticulation and/or phonetic conditioning, but no one perceives this difference in certain tasks, can we really call this a mismatch between production and perception, or is it simply a case where imperceptibility is one completely expected outcome, given the source of the production difference? These questions are reminiscent of Evanini's (2009) discussion of the English low back merger, discussed in Section 3.4.1.1 above.

This also calls into question what it means to produce a difference between two vowels in a case where phonetic/coarticulatory conditioning is at play. Can Speaker 4 be said to be "producing a difference" in the same way as Speaker 3 (**Figures 3.4-3.5**), or does that fact that one might be easily perceptible and socially meaningful, while the other might be imperceptible or perceptible under more limited conditions, mean that they are fundamentally different despite their similar directionality?

Following from these considerations of what it means for two vowels to be "different," the case of [a]/[ɒ]$_A$ in Gavaṛ also prompts us to reconsider what it means for vowels to be "merged" in production and perception. Given that the existence of a synchronic coarticulatory process means that a full merger of [a]/[ɒ]$_A$ in production (i.e., no consistently measurable phonetic difference between the two vowel classes) might never be expected even if there is a change in progress away from the dialect system and toward a more merged state, should we redefine "merger" in a way that will delineate perceptible and potentially socially meaningful differences in production from those that are not perceptible and/or not socially meaningful? How can/should we account for individual differences in perception and other circumstances that affect perception in defining merger in production?

The present study provides significant potential for further investigation of such questions, given that phonological and social factors are simultaneously involved in the examined phenomena and planned tasks. While different perceptual behaviors as manifested in different task types undoubtedly have implications for our understanding of concepts like "merger, "difference" and "perception-production mismatch," the additional factors at play in the present study – namely, the simultaneous involvement of socially salient allophony and synchronic coarticulation – force us to consider and define these concepts even more carefully. The examination of individual variation will add further nuance to these considerations, shedding light on the question of what it would mean for vowels to be "merged" in production and perception at the individual and community levels.

3.5 Overview of 2025 study methods and procedures

3.5.1 Methods

The 2025 study consisted of three components:

1) A speech production task, in which participants were recorded reading a list of words that contained the relevant vowel sounds (detailed methods and results presented in Chapter 2)
2) An AX discrimination task, which investigated participants' discrimination of differences between pairs of vowels that are found in the local dialect's vowel system
3) A metalinguistic perception task that involved categorizing tokens (many of which consisted of the same stimuli from the AX task) according to whether or not they are characteristic of Gavaṛ's dialect (this task is referred to as the "dialect categorization task" throughout this dissertation)

The AX task addresses the first research question (see Section 3.2 above), while the dialect categorization task addresses the second research question. Comparison of the two tasks addresses the third question, while comparison of the perception results with the production results addresses the fourth question.

The first perception task tested participants' discrimination of differences for two vowel pairs – [a]/[ɒ]$_A$ and [ɔ]/[ɒ]$_O$ – without any explicit social information or prompting. The second

perception task asked participants to categorize tokens from the [a]/[ɒ]$_A$ and [ɔ]/[ɒ]$_O$ continua one at a time according to their dialect association (Gavaṛ vs. not Gavaṛ). Both perception tasks are similar to tasks that have been used in previous sociophonetic perception research (see Thomas [2002] for an overview) and are very similar to the two types of perception tasks used in Butcher's (2022) study of several vowel mergers. The stimuli for both tasks were created from the speech of a single previous participant of the 2022 sociolinguistic interviews, who gave consent for portions of his interview to be presented as stimuli to participants of future studies. Tasks in which participants must associate different stimuli with various social characteristics have been used in many previous sociolinguistic perception studies (e.g., Casasanto 2010; Graff et al. 1986; Kerswill 2001; Morris 2010; Squires 2013). Graff et al. (1986) used a design in which listeners were asked to classify stimuli that were resynthesized from a single voice, as well as stimuli representing different dialects produced by the same bi-dialectal speakers, and in both cases the different guises were often classified differently by the participants. A subset of the stimuli in Kerswill (2001) were also taken from a single bi-dialectal speaker, and participants rated these stimuli as very different in their dialect association despite their having been pronounced by the same voice. Given these precedents, it was reasonable to assume that the use of a single voice for all the stimuli in the present study would not be a hindrance to participants' categorization of stimuli, as long as adequate instructions were provided.

The results of the two tasks will allow for the comparison of perception of the same vowel stimuli, representing two vowel pairs with differing social salience and phonological characteristics, in contexts both with and without social framing. This comparison will shed light on the influence of both social and phonological factors on perception. Together with the data from the speech production task, these results will also allow for an investigation of individual differences and the links between speech production and various perceptual behaviors, as described above.

3.5.2 Procedures

The production task and two perception tasks were carried out in Gavaṛ in June of 2025 with the help of the same local research assistant who conducted the sociolinguistic interviews in Gavaṛ. The first 7 participants completed the study in a different room than the remaining participants, due to constraints on the availability of different rooms at the Geghark^{h}unik^h Regional Library. There were also 2 participants (Participants 39 and 40) who completed the

study in a quiet room at a private home, due to difficulties in accommodating these participants' work schedules. In total, 51 participants completed the study (see Section 2.4.1 for more information about the participants). All three tasks were completed on the same day in order to maximize participant retention, with the production task preceding the perception tasks so that participants' productions would not be influenced by the perception stimuli. Participants signed an informed consent form before beginning the study. After completing all three tasks, participants completed a brief language background and demographic questionnaire, as well as an additional consent form that allowed them to give permission for various future uses of their data. This questionnaire can be found in Appendix B, while the consent forms can be found in Appendix C. The experiment materials and consent forms were reviewed by the University of Michigan IRB (HUM00267270), and the study was determined to be exempt from ongoing IRB review. On average, the entire process of participating in the study took one hour per participant. Each participant was compensated with 20 USD.

All participants interacted with both me and the research assistant. During the recording of the word list, only the research assistant remained in the room with the participant, both so that background noise would be minimized and so that participants would be less likely to alter their speech in the presence of a non-local person. Both the research assistant and I were present in the room during the perception tasks and the signing of informed consent forms. The research assistant read instructions to the participants before the recording of the word list and before each of the perception tasks. The text (in Armenian and in English translation) of these instructions can be found in Appendix D. In order to avoid creating an overly formal atmosphere, and with consideration for Armenia's diglossic environment, the research assistant relayed the instructions in a slightly more colloquial manner than what is represented in the text (for example, using the colloquial 3rd person singular auxiliary verb *a*, rather than SEA *ɛ*). The instructions included the same basic information about the tasks that was displayed on the screen (see Chapters 4 and 5 for figures showing the instruction screens) and emphasized that there were no correct or incorrect answers in the tasks, that the experiment simply sought to gain insights into how participants perceive different words and sounds, and that they should make decisions based on their initial reactions even if they sometimes felt like they were guessing.

All three parts of the study were completed using a laptop computer. The word list was displayed using Microsoft PowerPoint slides, with words being presented one at a time on

separate slides. The research assistant advanced the slides while the participant read the word list. Both perception tasks were implemented in PsychoPy (Peirce et al. 2019). Though an online version of the experiment was used for piloting, the final version conducted in Gavaṛ was run offline in order to avoid disruptions caused by unreliable internet connectivity. Participants heard the stimuli through AKG K271 MKII over-the-ear headphones and provided responses for both tasks by pressing the 0 and 1 keys. The computer volume was set to the maximum level before each participant began the experiment, and participants were instructed before the AX task to adjust the volume to a comfortable level if necessary during the practice portion of that task. Throughout the task (e.g., after each timed break and between the two tasks), the experiment included extra screens that instructed participants to press the spacebar to continue, so that participants would not accidentally start any portion of the experiment before they were ready. After the second perception task, participants were prompted with the screen shown in **Figure 3.7** and asked whether they thought they know the person whose voice was used for the stimuli. It was hypothesized that participants' impressions about whether or not they recognized the talker could influence their responses (particularly if they believed the talker to be a specific person whom they believed to use or not use dialect features in their speech). However, only 5 participants answered that they believed that they recognized the talker, and performance on the second perception task was very mixed among these participants, indicating that this may not have been an important factor in explaining participants' response patterns.

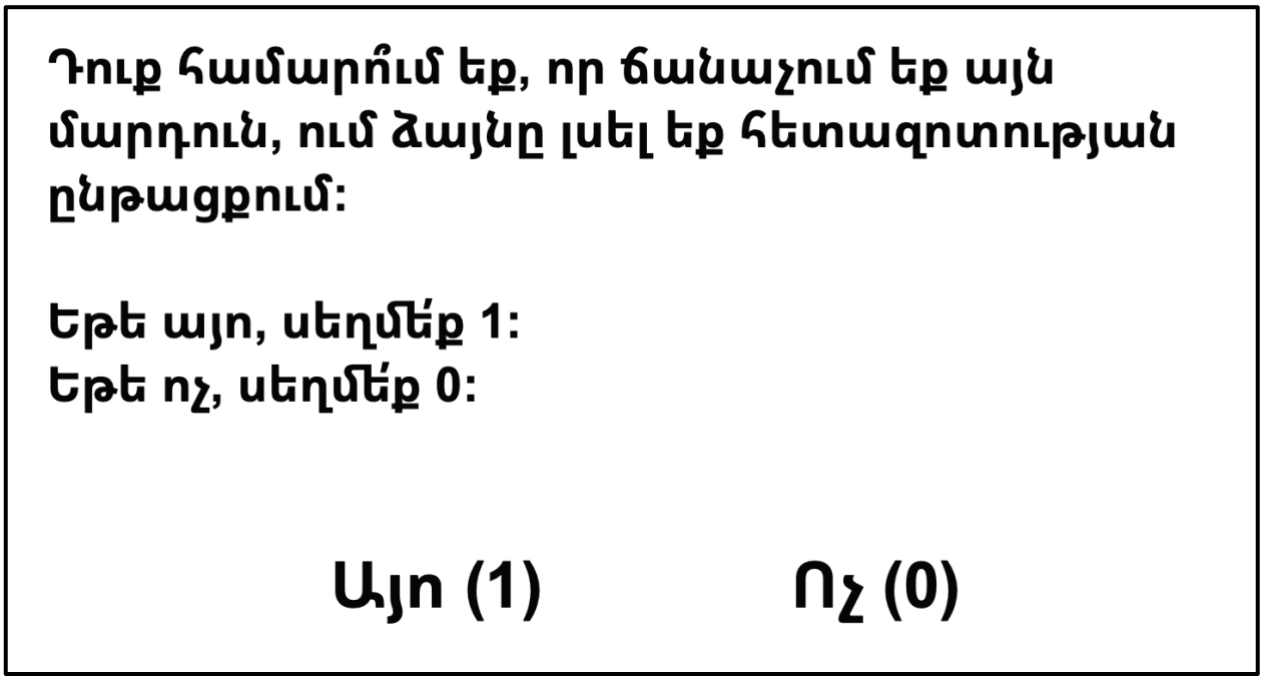


Figure 3.7: Screen shown at the end of the experiment, prompting participants to answer whether or not they thought they knew the person whose voice was used for the stimuli. The screen reads, "Do you think that you know the person whose voice you heard during the study? If yes, press 1. If no, press 0."

Three participants (Participants 1, 16, and 42) accidentally exited the experiment by pressing the "escape" key, which was located near the 1 key. Participant 1 accidentally exited the experiment during the second block of the AX task and subsequently restarted the experiment from the beginning of this block. As such, the participant responded twice to some trials of this block (i.e., the trials up to the point at which they accidentally exited). The participant's original responses to these trials were used in analyses, rather than the responses from the participant's second time hearing these trials. All subsequent participants were warned to be careful not to press the "escape" key, but a similar issue nevertheless occurred twice more. Participant 16 accidentally exited the experiment after the training block of the AX task. This participant's perception results (other than their responses to the AX training block) were unfortunately entirely lost, as entering an identical participant ID in PsychoPy for their first and second attempts caused the responses for the second attempt to not be saved. Participant 42 accidentally exited the experiment during the 3rd block of the AX task, and this participant's answers from their first and second attempts were used in analyses in the same manner as Participant 1's responses.

Chapter 4 AX Discrimination Task: Evidence for Influence of Phonological Characteristics on Discrimination

4.1 Chapter summary

This chapter presents the goals, expectations, methods, and results for the AX task. The chapter begins by summarizing the specific ways that the task investigates participants' discrimination of different pairs of vowels found in Gavaṛ's dialect, along with expectations about the results of the task. The task design and stimuli are then discussed, followed by the methods used for analyzing the data, which included calculation of various metrics of performance on the task, as well as statistical methods predicting responses to the task (mixed-effects linear and logistic regression models). The presentation and discussion of the results focuses on differences and similarities between the two vowel pairs under investigation ([a]/[ɒ]$_A$ and [ɔ]/[ɒ]$_O$), highlighting how the results compare with expectations and considering various explanations for observed patterns. The main finding was that discrimination of [a]/[ɒ]$_A$ stimuli was unexpectedly poor considering that the acoustic differences in the stimuli were much larger for this vowel pair than for [ɔ]/[ɒ]$_O$. The most likely explanation for this discrepancy is the allophonic status of [a]/[ɒ]$_A$, as previous studies have found that listeners generally have more difficulty discriminating allophonic differences than contrastive ones. Discrimination of this vowel pair was especially poor near the end of the vowel continuum representing small differences from the Standard Eastern Armenian (SEA) variant, which could suggest that participants perceived differences in this area of the vowel continuum as coarticulatory effects similar to those found in Yerevan Armenian. Overall, the results also suggested that older people and people without higher education were more successful at discriminating the stimuli.

This chapter also compares different possible ways of analyzing the data, discussing differences in the results that stem from different methodological decisions, and joining previous authors in stressing the need for methodological transparency and avoidance of arbitrary decisions.

4.2 Goals and expectations of the AX task

As described in Chapter 3, the [a]/[ɒ]$_A$ allophony in Gavaṛ may have originated from enhancement of one or more coarticulatory processes, un-enhanced versions of which might be seen in the vowel spaces of those speakers who adhere less closely to the local dialect's vowel system. Alternatively, the latter speakers may simply be more advanced representatives of a sound change in progress. In principle, both accounts can coexist, as there could be an ongoing shift toward a more merged state that would nonetheless never reach completion due to the synchronic coarticulatory processes of the sort found by Seyfarth and Garellek (2018). In any case, the 2022 data (presented in Chapter 2 and further discussed in Chapter 3) showed that residents of Gavaṛ vary vastly in the magnitude of the differences that they produce between [a] and [ɒ]$_A$. Accordingly, the research question investigated through the AX task asks about the magnitude and type of acoustic difference necessary for residents of Gavaṛ to perceive these two vowels as distinct. In order to see if the particular characteristics of the [a]/[ɒ]$_A$ comparison (being completely allophonic, having potential roots in coarticulation, being socially salient) affect the discrimination of differences between the two vowels, the task also investigates discrimination of differences between [ɔ] and [ɒ]$_O$, which represent distinct phonemes despite exhibiting contextually conditioned distributions in some subsets of vocabulary (see Katvalyan 2016: 89-91). Thus, the core question investigated through the AX task is whether/how discrimination of acoustic differences varies across these two pairs of vowels, with consideration for their phonological and social characteristics. The potential role of listeners' demographic characteristics in predicting their performance on the task is also explored in this chapter.

For both the [a]/[ɒ]$_A$ and [ɔ]/[ɒ]$_O$ comparisons, significant gender patterns were found in the 2022 sociolinguistic interview data, as discussed in Chapter 2. However, unlike the data for [a]/[ɒ]$_A$, the data for [ɔ]/[ɒ]$_O$ did not reveal a pattern in which all participants maintained at least a small difference between the two vowels with the same directionality as the larger differences maintained by other participants (for [ɔ]/[ɒ]$_O$, the differences in production were overall smaller, with a larger portion (around half) of participants having Pillai scores deemed merged by the method of Stanley and Sneller [2023]). There is also no evidence that [ɔ]/[ɒ]$_O$ (which is contrastive) has roots in coarticulation, and it is likely less socially salient. Given these differences, it was hypothesized that discrimination of differences between [a]/[ɒ]$_A$ and [ɔ]/[ɒ]$_O$ would differ in some way.

As described in Section 4.3, the AX task tested discrimination both of differences of varying magnitudes and of differences of a fixed magnitude at different locations along continua ranging between the two endpoints of each vowel pair.

Regarding the investigation of differences of varying magnitudes, it is possible that smaller differences on the [ɔ]/[ɒ]$_O$ continuum might be perceived, since these two vowels are contrastive according to Katvalyan (2016), while small differences in [a]/[ɒ]$_A$ might not be perceived due to this pair's allophonic nature and potential coarticulatory roots. On the other hand, smaller differences in [a]/[ɒ]$_A$ could be perceived because of its social salience, which could have some effect on how acoustically small differences are perceived. As will be described in Section 4.3.2.1, the acoustic differences presented in the stimuli were generally larger for [a]/[ɒ]$_A$, meaning we would expect discrimination of this pair to be better if the phonological characteristics and social salience of the different vowels have no effect. Likewise, if acoustic size of differences in stimuli is the main factor influencing accuracy of discrimination, we would expect individual words (within each vowel pair) to be discriminated with differing accuracy in accordance with the size of their acoustic differences.

Regarding the investigation of differences of a fixed magnitude at various locations along the continua, it is possible that listeners will show a clearer peak in perceptual accuracy at some point along the [ɔ]/[ɒ]$_O$ continuum in comparison with the [a]/[ɒ]$_A$ continuum, again due to the contrastive status of this difference. When two sounds are contrastive, we can generally expect that listeners will switch from hearing one to hearing the other at some point along a continuum, and thus that they will be more accurate at discriminating differences near this crossover point than at other points along the continuum (see Section 4.3.1.1; see also discussion of categorical perception in Section 3.4.2). We might expect to see such a peak in the data for [ɔ]/[ɒ]$_O$, while for the allophonic [a]/[ɒ]$_A$ comparison, listeners might show similar accuracy along the entire continuum. On the other hand, if we consider the findings of Seyfarth and Garellek (2018), which suggest that small differences from [a] in the direction of [ɒ]$_A$ are a coarticulatory effect, we might expect participants to have more difficulty discriminating small differences from the [a] endpoint of this continuum. Though the task is not a direct examination of perceptual compensation for coarticulation, small differences from this endpoint can be considered the most analogous to the differences identified by Seyfarth and Garellek (2018), based on the magnitude of the differences and where they are located in the vowel space.

As noted in Chapter 3, the analyses of the AX task presented in this chapter include some exploration of relationships between performance on the AX task and production of the relevant vowels. However, since other aspects of relationships between perception and production are explored in the analyses presented in Chapters 5 and 6, the main discussion of this issue will be presented in Chapter 6.

4.3 AX task – stimuli, task structure, and data collection

4.3.1 Types of stimuli

4.3.1.1 Test stimuli

The test stimuli, representing the [a]/[ɒ]$_A$ and [ɔ]/[ɒ]$_O$ continua, were words with different predicted pronunciations in SEA and Gavaṛ's dialect, such as [pat]/[pɒt] for [a]/[ɒ]$_A$ and [t͡sɔt͡s^h]/[t͡sɒt͡s^h] for [ɔ]/[ɒ]$_O$. These pronunciations represented the endpoints of each continuum, while the remaining steps ranged between these endpoints. Importantly, though [ɔ]/[ɒ]$_O$ is a phonemic contrast according to Katvalyan (2016), the words used as stimuli did not represent minimal pairs, but rather represented different pronunciations of the same word, just like the stimuli representing [a]/[ɒ]$_A$. (As noted in Section 2.4.2 in the discussion of the production study word lists, the minimal pairs listed by Katvalyan (2016) for [ɔ]/[ɒ]$_O$ generally consisted of words with differing stylistic associations [e.g., a given pair might consist of one dialectally neutral word and one strongly associated with Gavaṛ's dialect]). The words used for the test stimuli are presented in **Table 4.1** in Section 4.3.2.1 below.

As noted in Section 4.2, the test trials for the [a]/[ɒ]$_A$ and [ɔ]/[ɒ]$_O$ vowel pairs were broken down into two groups. The trials consisted both of those intended to test the discrimination of differences of varying magnitudes, and those intended to test the discrimination of differences of a fixed magnitude at different locations along the continua. It was necessary to explore both of these avenues because of the relative lack of previous studies probing the perception of allophonic vowel variation with stimuli that vary along continua, which made it difficult to precisely predict how listeners might behave in the present study. Amengual and Chamorro (2015) conducted an AX task involving stimuli whose vowels varied along continua, testing only intervals of a fixed magnitude at different locations along the continua. They were examining vowel differences that are phonemic in Galician (but not in Spanish), with participants who were either Galician-dominant or Spanish-dominant Galician-Spanish

bilinguals. Though the Galician-dominant participants showed strong peaks in perceptual accuracy at a certain point along each continuum, indicating the presence of a phonemic category boundary, the Spanish-dominant participants showed no or much weaker peaks in perceptual accuracy. This suggests that listeners perceiving vowel continua whose endpoints do not represent separate phonemes for them may show similar perceptual accuracy at all points along the continua. This finding aligns not only with the broader literature on categorical perception (see Section 3.4.2), but also with other studies that examine the effects of linguistic experience on discrimination of other types of sounds (see, e.g., Miyawaki et al. [1975] regarding discrimination of consonant continua).

Accordingly, since the AX task of the present study involves one entirely allophonic vowel pair ([a]/[ɒ]$_A$) and one that sometimes shows a phonetically conditioned distribution despite the existence of some minimal pairs ([ɔ]/[ɒ]$_O$), presenting listeners only with intervals of a fixed magnitude at different locations along the continua could have produced relatively uninformative results. Though it was possible that the salience of the [a]/[ɒ]$_A$ allophony and the contrastive nature of the [ɔ]/[ɒ]$_O$ difference could have led listeners to nevertheless exhibit category boundaries in perception, the possibility that they would exhibit no such boundaries meant that an additional method was needed in order to probe their discrimination of differences between [a]/[ɒ]$_A$ and [ɔ]/[ɒ]$_O$. The subset of the AX trials which presented differences of varying magnitudes was intended to determine the distance from the canonical SEA vowel ([a] or [ɔ]) that was needed in order for each listener to start perceiving a difference. This portion of the AX task thus included trials that compared the [a] and [ɔ] endpoints (both step 7 in their respective continua) to every other step along their own continuum (i.e., step 7 – step 6, step 7 – step 5, step 7 – step 4, etc.).

4.3.1.2 Training, filler, and control stimuli

The training portion of this task used stimuli that are predicted (according to previous dialect descriptions such as Katvalyan [2016, 2009]) to contain the vowel [æ] in Gavaṛ's dialect and the vowel [a] in SEA, along with stimuli containing the vowel [u] with various F2 manipulations (see Sections 4.3.2.1 and 4.3.2.2 for more details about the creation of these stimuli). The [a]/[æ] stimuli, which ranged along a continuum, were also used as fillers in the test portion. In the case of both groups of training stimuli, participants heard differences of varying magnitudes, as well as "same" trials. For the [æ] stimuli, "different" trials consisted either of

endpoints of the continua, or of differences of 3 steps, which were determined to present moderate difficulty of discrimination. Piloting confirmed that these differences were discriminable for participants some (but not all) of the time. For the [u] stimuli, participants heard both "same" and "different" trials, with "different" trials consisting of differences that were relatively easy and relatively difficult to perceive, as confirmed by piloting. In general, the AX task was piloted both by people from Gavaṛ and by speakers of other varieties of Armenian, as the goal of piloting this task was to determine the size of acoustic differences that would present an appropriate level of difficulty.

The group of words where [æ] is predicted in the dialect was suitable for use as training and filler stimuli for several reasons. First, the 2022 interviews demonstrated that there is some metalinguistic awareness about this vowel, meaning that we could expect the participants of this task to notice and attend to it. Additionally, the vowel [æ] is an independent phoneme according to Katvalyan (2016: 83-85), meaning that participants might more reliably perceive it as different from [a], in comparison with allophonic differences. Moreover, this group of words is less useful for the experimental tasks because the [æ] vowel and many of the words containing it[23] are not specific to Gavaṛ's dialect. Katvalyan (2016: 55-56; 2025: 34) writes that [æ] and other fronted vowels are common across many dialects of Armenian, and Katvalyan (2025: 34) and Mkrtch^{h}yan (2016) mention this vowel occurring in other dialects of Gegharkʰunikʰ province. Because of the potential lack of geographic specificity associated with the [æ] vowel, this group of words would have been unsuitable for the second perception task (dialect categorization task), in which listeners had to classify pronunciations according to their dialect association. As the results of the two perception tasks were also intended to be compared with one another, this group of words was also not suitable for use as one of the test conditions of the AX task. The suitability of this group of words for training purposes was augmented by the fact that the [a]/[ae] contrast would likely be less obvious to participants than phonemic contrasts found in all varieties of Armenian (such as [u] vs. [i]), which would help to prepare participants for the subtle vowel differences found in the main portion of the task.

Since the [a]/[æ] distinction is nevertheless phonemic in Gavaṛ's dialect, the training portion also included words containing [u] with various F2 manipulations. For example,

[23] According to Katvalyan (2016: 83-85; 2025: 34), [æ] is largely found in loanwords, and I have witnessed several of Katvalyan's listed loanwords with [æ] being used by speakers of other varieties of Armenian.

participants heard a word containing an unmodified [u] together with the same word containing an [u] with a raised F2, resulting in the impression of fronting of the vowel. Several words containing [u] were presented with different degrees of fronting, in order to help participants become accustomed to attempting to discriminate non-phonemic vowel differences of varying magnitudes.

In addition to the [a]/[ɒ]$_A$ and [ɔ]/[ɒ]$_O$ test continua and the [a]/[æ] filler continuum, the main portion of the task included control tokens representing several obvious phonemic vowel contrasts, which were used to verify that participants were performing the task correctly. For example, if participants heard the words [pɛt] (‘commander’) and [put] (‘speck’) and selected that the pronunciations are different, then it would be clear that they had understood the task. The words used as controls are presented in Section 4.3.2.2. Though it was possible that the inclusion of such pairs could have caused participants to attend less closely to subtle vowel differences, pilot data showed that all participants were discriminating non-phonemic differences, which suggested that the instructions and training portion were sufficient for demonstrating the nature of the task.

4.3.2 Creation of stimuli

As noted above, all stimuli were resynthesized from the speech of one participant of the 2022 sociolinguistic interviews (Speaker 3, whose vowel space is shown in Section 3.3, **Figure 3.4**), who gave consent for his interview data to be used as stimuli in future studies. This participant was chosen to become the talker of the perception tasks because his vowel space showed relatively close adherence to the previously described dialect vowel system, such that his productions could serve as the basis for [a]/[ɒ]$_A$, [ɔ]/[ɒ]$_O$, and [a]/[æ] continua. His productions of the differences between [a]/[ɒ]$_A$ and [ɔ]/[ɒ]$_O$ were also among the most distinct among the 2022 participants. Before finalizing the choice of the talker, it was also confirmed that the given participant’s productions from the sociolinguistic interview data could provide all other control and filler items for both perception tasks.

It was decided that all stimuli should be drawn from the speech of one previous participant in order to minimize the number of previous participants who would not be able to participate in the 2025 study. Having participants hear only one voice throughout both perception tasks was also preferable for the purpose of controlling for extraneous factors that could influence perception if different voices were used. Additionally, because the second perception

task involved classifying pronunciations by their perceived dialect association, if several voices had been used throughout the tasks, then participants might have been confused as to why they were being asked to associate tokens obviously pronounced by only one of the several voices with different dialects. In other words, participants might have had trouble understanding, for example, that they should associate [pat] and [pɒt] with different dialects, if they were hearing these words produced by the same talker who was only one of several from whom stimuli had been drawn. On the contrary, if a single voice were used for all stimuli, and if participants were instructed that they will only hear one person's voice and thus should focus on differentiating and grouping the pronunciations themselves, then it should be more clear that the same voice's [pat] and [pɒt] can be associated with different dialects.

The intensity of all stimuli in both portions of the task was normalized to 71 decibels using Matt Winn's "Scale intensity of all sounds in a Directory" Praat script (Winn 2023). This amplitude was chosen because it did not result in any files being clipped.

4.3.2.1 Resynthesized vowel continua (test and training/filler)

The [a]/[ɒ]$_{A}$, [ɔ]/[ɒ]$_{O}$, and [a]/[æ] vowel continua were resynthesized using a Praat script created by Matt Winn (Winn 2023). Although the *Akustyk* package for Praat has been used for resynthesizing vowel continua in many previous studies (e.g., Amengual & Chamorro 2015; D'Onofrio 2021; Loakes et al. 2014), it was no longer available for download as of March 2024, and Winn's script provided a suitable alternative. Tokens that served as the endpoints of each continuum were selected based on distinctiveness of the vowels from one another and on adequacy of recording quality.

With consideration both for the amount of different lexical items that could feasibly be included in an experiment of a reasonable length, and for the number of distinct word pairs that could be obtained from the talker's previous data, two words were selected to represent each test continuum ([a]/[ɒ]$_{A}$ and [ɔ]/[ɒ]$_{O}$) in the AX task, while four words were selected to represent the training/filler continuum ([a]/[æ]). These words, as well as the corresponding words from which the other endpoint of each continuum was drawn, are summarized below in **Table 4.1**. In choosing these latter words, an attempt was made to hold the preceding and following places of articulation as constant as possible. For example, the [a]/[ɒ]$_{A}$ continuum for the word [pat]/[pɒt] was synthesized by using a [pɒt] token as the basis for the [ɒ]$_{A}$ end of the continuum, and the vowel from a [bad] token as the basis for the [a] end of the continuum. As noted in Section 2.4.2,

holding both the preceding and following places of articulation constant was not possible in the case of the [ɔ]/[ɒ]$_O$ continuum. Thus, for these words, only the following place of articulation was held constant. Both places of articulation were mostly held constant in the case of the [a]/[æ] continua. However, this requirement was relaxed for the word [naft^h]/[næft^h] (‘oil’), as a continuum with more distinct endpoints was obtained by pairing it with a word with a different following place of articulation (see **Table 4.1** below). It was also relaxed in the case of the word [t^{h}at^h]/[t^{h}æt^h] (‘paw’), where the first vowel from the word [ædæt^h] (‘custom’) provided the best counterpart to the vowel from [t^{h}at^h].

Table 4.1: Test and training/filler stimuli of the AX task. The “word with desired consonantal context” and “word providing other endpoint” columns represent the actual tokens obtained from the talker’s sociolinguistic interview data, while the “resulting continuum” column represents the continua endpoints that were obtained after resynthesis.

vowel pair	**word with desired consonantal context**	**word providing other endpoint**	**resulting continuum**
[a]/[ɒ]$_A$	[pɒt] (‘wall’)	[bad] (‘duck’)	[pat]/[pɒt]
	[mɒs] (‘part’)	[bad] (‘duck’)	[mas]/[mɒs]
[ɔ]/[ɒ]$_O$	[p^{h}ɒs] (‘hole’)	[ɔt^h] (‘air’)	[p^{h}ɔs]/[p^{h}ɒs]
	[t͡sɒt͡s^h] (‘bosom, embrace’)	[ɔt^h] (‘air’)	[t͡sɔt͡s^h]/[t͡sɒt͡s^h]
[a]/[æ]	[d͡zæk^h] (‘baby animal’)	[d͡zaχ] (‘left’)	[d͡zak^h]/[d͡zæk^h]
	[dærd] (‘sorrow’)	[darə] (‘bitter’)	[darɖ]/[dærd]
	[næft^h] (‘oil’)	[das] (‘lesson’)	[naft^h]/[næft^h]
	[t^{h}at^h] (‘paw’)	[ædæt^h] (‘custom’) (first vowel)	[t^{h}at^h]/[t^{h}æt^h]

Before the creation of the continua, vowel durations were modified in the words that would become the test stimuli (two words per vowel pair for the AX task, plus an additional third word per vowel pair for the dialect categorization task; see Chapter 5). The decision to modify vowel length was based on several factors, including the impression that significantly shorter

vowels sounded less discriminable even in cases where formant differences were approximately equal. Preliminary pilot data seemed to confirm this impression, and in general, it was decided that stark differences in vowel length could be jarring or distracting for participants.

Some previous studies that involve vowel stimuli varying along continua (e.g., Fridland & Kendall 2012; Kendall & Fridland 2012) hold vowel duration constant across different words. However, a different strategy was adopted in the present study, in part due to the observation that some of the variation in vowel length found in the words used for the test stimuli was likely a result of coarticulatory influence of the surrounding consonants. For example, the vowel in the word [χɔz]/[χɒz] ('pig'), used as the additional word in the dialect categorization task to represent the [ɔ]/[ɒ]$_{O}$ vowel pair (see Chapter 5), was substantially longer than the vowel in all the other words being used as test stimuli, possibly because of the influence of the voiced fricative following the vowel. Completely eliminating variation in vowel length could have thus made some of the stimuli sound unnatural (e.g., if the vowel were much shorter than the one that listeners would expect in a given consonantal context), potentially impacting the results. It was thus decided that the vowel durations should be brought closer together, but not made completely identical, such that some of the natural coarticulatory variation in vowel length would be maintained. It was also decided that the vowel durations in all test words should be altered, such that any unnaturalness resulting from altering vowel duration would affect all words approximately equally. Similar methods were implemented by Beddor et al. (2024), who altered all stimuli vowel durations without making them identical.

Vowel durations were altered for all test words representing the [a]/[ɒ]$_{A}$ and [ɔ]/[ɒ]$_{O}$ vowel pairs, but not for the [a]/[æ] filler stimuli. Among the words used to create the test stimuli, vowel duration was only modified in the words whose consonantal context would ultimately be used in the final continua. Modifying duration in the other set of words was unnecessary because these words only provided information related to formant values and trajectories, while the vowel duration of the resynthesized tokens came from the words in the "word with desired consonantal context" column of **Table 4.1**. The vowel durations in test words were modified using the PSOLA technique, implemented as Praat's *manipulate* function (Boersma & Weenink 2021). The durations were modified such that they all fell within one standard deviation of the mean of the original durations. The original vowel durations and new durations for each word are shown below in **Table 4.2**.

Table 4.2: Original vowel durations and modified vowel durations in milliseconds for the six test words used in the AX and dialect categorization tasks (the words [marʳ^{h}]/[mɒrʳ^{h}] ('person') and [χɔz]/[χɒz] ('pig') were used only in the dialect categorization task).

word	original vowel duration (ms)	new vowel duration (ms)	difference between original and new durations (ms)
[pat]/[pɒt] ('wall')	242	255	13
[mas]/[mɒs] ('part')	232	250	18
[marʳ^{h}]/[mɒrʳ^{h}] ('person')	240	253	13
[p^{h}ɔs]/[p^{h}ɒs] ('hole')	235	244	9
[t͡sɔt͡s^{h}]/[t͡sɒt͡s^{h}] ('bosom, embrace')	201	232	31
[χɔz]/[χɒz] ('pig')	286	259	-27

In an effort to further normalize the duration of the stimuli, the silence before the beginning and after the end of each word was cut to 100 ms. The preceding 100 ms was measured from the first visible disturbance in the waveform/spectrogram related to the word's initial consonant, while the following 100 ms was measured from the end of the stop burst or frication of the final consonant, as judged auditorily and visually based on the waveform/spectrogram.

After the modification of vowel durations and removal of silence, 7-step continua were created using Winn's Praat script. Only F1 and F2 were modified (no F3), and the Bark method of interpolation was selected, such that the steps would be equally spaced along the Bark scale. In determining the number of steps to be created, and the pairs of steps that were to be compared with one another, the general goal was to create pairs of steps that would be neither too difficult nor too easy to discriminate, such that participants were unlikely to perform either at floor or at ceiling. Continua of 7 steps were determined to be suitable for the experiment structure after piloting confirmed that 2-step or 3-step differences (depending on the vowel pair) were sometimes discriminable for all participants, while 1-step differences were much more difficult to discriminate.

The "maximum formant value" setting of Winn's Praat script was adjusted via trial and error until natural-sounding stimuli were obtained (maximum formant settings that were too high

or too low resulted in non-speech-like sounds). The value used for each word in the AX task is listed in **Table 4.3**. As noted above, only F1 and F2 were modified in the creation of these continua (F3 was not modified), such that this maximum formant setting represents the formant below which the script identified F1 and F2.

The mean formant values in Hz and Bark of each of the 7 steps of the continua for the [a]/[ɒ]$_A$ and [ɔ]/[ɒ]$_O$ vowel pairs in the AX task are presented in **Table 4.4** and **4.5**. As the tables show, the values differ between the two words representing each vowel pair ([pat]/[pɒt] vs. [mas]/[mɒs] and [p^{h}ɔs]/[p^{h}ɒs] vs. [t͡sɔt͡s^h]/[t͡sɒt͡s^h]). This is because the endpoint formant values are based on the talker's natural tokens of each word. **Table 4.6** shows the difference in Bark for F1 and F2 between consecutive steps for each word. All Bark values were obtained using the *barktools* package (Stanley 2020).

As **Table 4.6** shows, when the formant values of the stimuli are converted into Bark units, the differences between steps of the [a]/[ɒ]$_A$ pair are larger both in F1 and F2 than the differences between steps of the [ɔ]/[ɒ]$_O$ pair. The consistent-size interval differences presented in the experiment were made more comparable by presenting 2-step intervals in the case of the former and 3-step intervals in the case of the latter. However, because the F2 differences between the steps of the [ɔ]/[ɒ]$_O$ pair were minimal to begin with, the presented F1 differences across the two vowel pairs are much more comparable than the F2 differences. The F1 differences presented in the experiment for the two words representing the [a]/[ɒ]$_A$ pair are 0.64 Bark and 0.78 Bark, while the F1 differences presented for the two words representing the [ɔ]/[ɒ]$_O$ pair are 0.57 Bark and 0.51 Bark. The F2 differences presented in the task for the two words representing the [a]/[ɒ]$_A$ pair are 0.58 Bark and 0.50 Bark, while the F2 differences presented for the two words representing the [ɔ]/[ɒ]$_O$ pair are both 0.14 Bark.

Table 4.3: Maximum formant value setting used in the creation of each continuum for the AX task with Winn's Praat script.

word	**maximum formant value setting**
[pat]/[pɒt]	2000 Hz
[mas]/[mɒs]	2000 Hz
[p^{h}ɔs]/[p^{h}ɒs]	2000 Hz

[t͡sɔt͡sʰ]/[t͡sɒt͡sʰ]	2000 Hz
[d͡zakʰ]/[d͡zækʰ]	2000 Hz
[dɑɾd]/[dæɾd]	2000 Hz
[naftʰ]/[næftʰ]	2200 Hz
[tʰatʰ]/[tʰætʰ]	2300 Hz

Table 4.4: Mean F1 values in Hz and Bark for each step of the [pat]/[pɒt], [mɑs]/[mɒs], [pʰɔs]/[pʰɒs], and [t͡sɔt͡sʰ]/[t͡sɒt͡sʰ] continua.

word	step 1		step 2		step 3		step 4		step 5		step 6		step 7	
	Hz	Bark	Hz	Bark	Hz	Bark	Hz	Bark	Hz	Bark	Hz	Bark	Hz	Bark
[pat]/[pɒt]	479	4.73	515	5.05	552	5.36	590	5.68	630	5.99	671	6.31	713	6.62
[mɑs]/[mɒs]	426	4.26	470	4.65	515	5.05	562	5.45	611	5.84	662	6.24	716	6.64
[pʰɔs]/[pʰɒs]	482	4.76	461	4.58	441	4.39	420	4.21	401	4.02	381	3.83	362	3.65
[t͡sɔt͡sʰ]/[t͡sɒt͡sʰ]	468	4.64	450	4.48	432	4.31	414	4.14	396	3.98	379	3.81	362	3.65

Table 4.5: Mean F2 values in Hz and Bark for each step of the [pat]/[pɒt], [mɑs]/[mɒs], [pʰɔs]/[pʰɒs], and [t͡sɔt͡sʰ]/[t͡sɒt͡sʰ] continua.

word	step 1		step 2		step 3		step 4		step 5		step 6		step 7	
	Hz	Bark	Hz	Bark	Hz	Bark	Hz	Bark	Hz	Bark	Hz	Bark	Hz	Bark
[pat]/[pɒt]	833	7.47	877	7.75	921	8.04	967	8.33	1015	8.62	1064	8.90	1115	9.19
[mɑs]/[mɒs]	872	7.72	909	7.97	948	8.21	988	8.45	1029	8.70	1071	8.94	1114	9.19
[pʰɔs]/[pʰɒs]	957	8.26	946	8.20	935	8.13	925	8.07	915	8.00	904	7.94	894	7.87
[t͡sɔt͡sʰ]/[t͡sɒt͡sʰ]	950	8.22	941	8.17	932	8.11	923	8.05	914	7.99	905	7.94	896	7.88

Table 4.6: F1 and F2 differences (Bark) between consecutive continuum steps for each test word in the AX task.

word	F1 step (Bark)	F2 step (Bark)
[pat]/[pɒt]	0.32	0.29
[mɑs]/[mɒs]	0.40	0.24

[pʰɔs]/[pʰɒs]	-0.19	-0.07
[t͡sɔt͡sʰ]/[t͡sɒt͡sʰ]	-0.17	-0.06

4.3.2.2 Other stimuli ([u] training stimuli and minimal pairs control condition)

The [u] training stimuli were created using the Praat plugin "Vocal Toolkit" (Corretge 2012-2024), which was used to change the mean F2 value in two words ([put] 'speck' and [d͡ʒuɾ] 'water') containing this vowel. In order to align with the other stimuli, these training stimuli were cut to contain 100 ms of silence both before the start of the initial consonant and after the end of the final consonant.

The minimal pairs used as a control condition in the test portion of the task are presented in **Table 4.7**. Vowel duration was normalized within each pair using Praat's *manipulate* function, with the two vowel durations in each pair being brought as close to their mean as possible. In the resulting stimuli, the differences between the vowel durations of the two members of each pair ranged from 0 to 6 ms. The duration of silence preceding the initial consonant and duration of silence following the final consonant were adjusted in the same manner described above for the test stimuli. The pitch contour was normalized within each pair by using Praat "Vocal Toolkit" to copy the pitch contour from one member of the pair to the other. Specifically, the pitch from [put] was copied to [pɛt], the pitch from [pʰiʁ] was copied to [pʰɔʁ], the pitch from [tɛs] was copied to [tas], and the pitch from [tʰutʰ] was copied to [tʰatʰ]. The member of each pair that provided the pitch contour was chosen based on whichever direction of copying resulted in more natural-sounding stimuli (one direction of copying generally resulted in stimuli that sounded distorted and noisy in comparison with the other direction of copying).

Table 4.7: Minimal pairs used as control stimuli in the AX task.

[pɛt] ('commander')	[put] ('speck')
[pʰiʁ] ('elephant')	[pʰɔʁ] ('money')
[tɛs] ('see' [imperative])	[tas] ('ten')
[tʰatʰ] ('paw')	[tʰutʰ] ('mulberry')

4.3.3 Task structure

4.3.3.1 Overall structure

There were 27 total training trials (10 "same" trials and 17 "different" trials), while the test portion of the AX task consisted of 392 total trials (152 "same" trials and 240 "different" trials). The training and test portions of the task thus consisted of 419 total trials. The test portion included two 2-minute breaks, such that it was divided into three portions of 131, 130, and 131 trials. These breaks were automatically timed, but when the timer ran out, a screen was displayed saying "press the spacebar to continue," so that participants could initiate the start of the next section themselves. All sections of the experiment displayed a progress bar that showed participants the portion of the current section remaining until the next break. The order of the trials in both portions was pseudo-randomized, so that any influence of the order of the trials would be the same for all participants. The trials in the test portion were arranged such that the first few and last few trials in each section were filler or control trials, so that the test trials would not be affected by factors such as participants being inattentive at the beginning of a new section or rushing through the last few trials of a section in order to reach the break. The AX task was not blocked by word or by vowel pair, since it was determined via piloting that the unblocked task was not excessively difficult. The AX task took approximately 25 minutes to complete.

The task instructions, which were displayed before the training portion and again before the test portion of the task, are shown below in **Figure 4.1**. The response screen is shown in **Figure 4.2**.

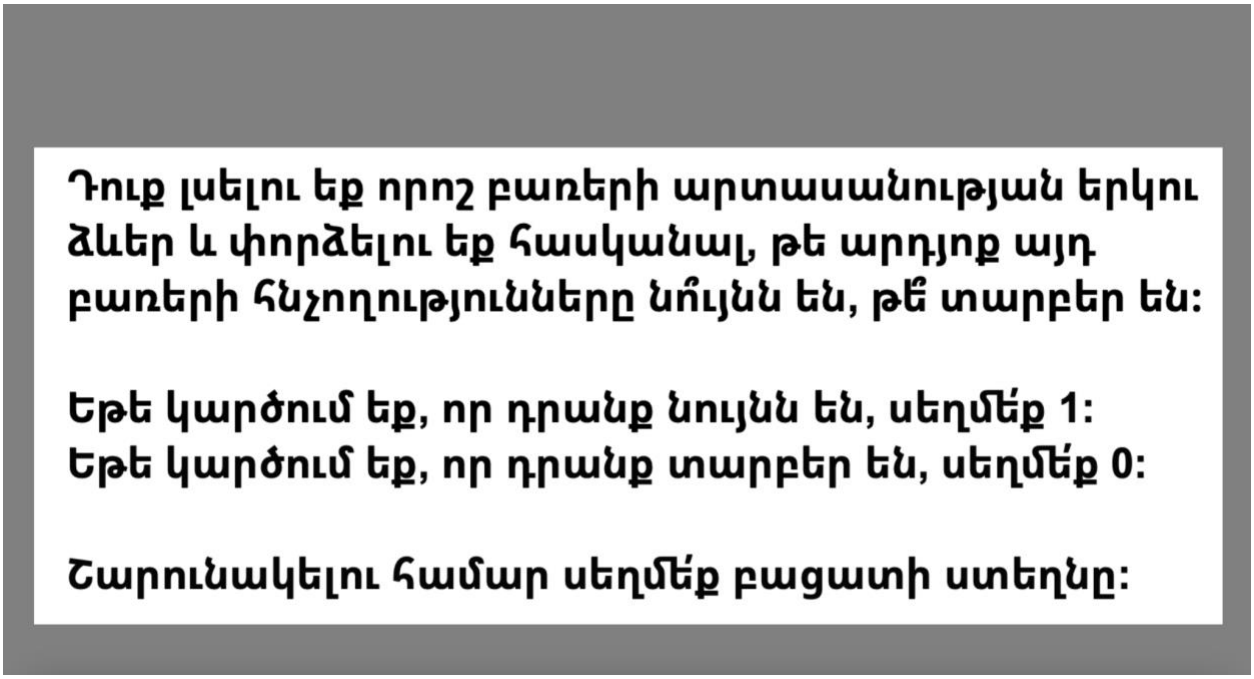


Figure 4.1: Instructions displayed at the beginning of the training and test portions of the AX task. The text reads, "You will hear two forms of pronunciation of some words and will try to understand if the ways that those words sound are the same or different. If you think they are the same, press 1. If you think they are different, press 0. Press the spacebar to continue."

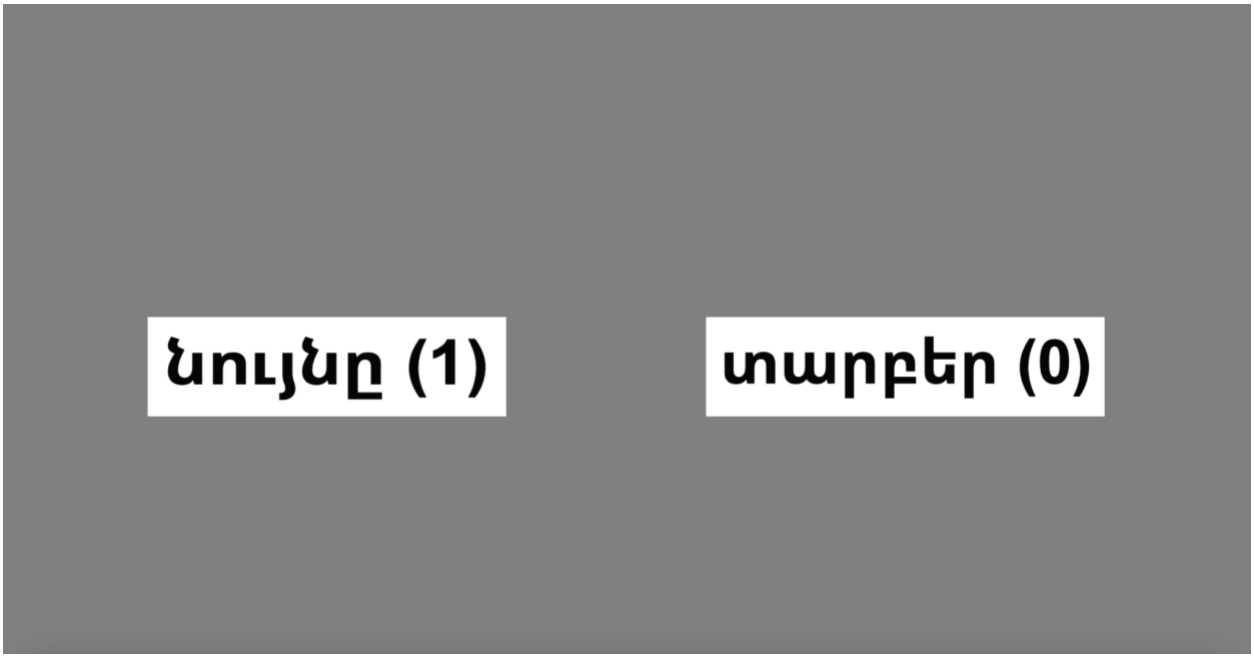


Figure 4.2: Response screen from the AX task. The left option reads "same (1)", while the right option reads "different (0)".

4.3.3.2 Structure of training portion

Table 4.8 below shows a complete breakdown of the training stimuli.[24] The table reflects the actual number of trials (i.e., most trials were not repeated during the training portion).

Table 4.8: Training stimuli for the AX task. Where relevant, the "word" column shows the predicted SEA and dialect pronunciations, which correspond to the two ends of the continuum for the given word. The order of the stimuli in the "pair" column reflects the order in which they were presented. The overall order shown below does not reflect the pseudo-randomized order used in the task.

word	pair	same or different?
[d͡zak^{h}]/[d͡zæk^{h}] ('baby animal')	step 1 – step 1	same
	step 4 – step 4	same
	step 7 – step 7	same
	step 1 – step 7	different (endpoints)
	step 7 – step 1	different (endpoints)
	step 1 – step 4	different (3 steps)
	step 6 – step 3	different (3 steps)
[daɾd]/[dæɾd] ('sorrow')	step 1 – step 1	same
	step 4 – step 4	same
	step 7 – step 7	same
	step 1 – step 7	different (endpoints)
	step 7 – step 1	different (endpoints)
	step 5 – step 2	different (3 steps)
	step 4 – step 7	different (3 steps)
[d͡ʒuɾ] ('water')	unmodified – unmodified	same

[24] Although the dialect pronunciation [d͡z^{h}æk^{h}] (with a breathy-voiced initial consonant) is listed by Katvalyan (2016), the presence or absence of breathiness is not relevant to the present study. Sub-optimal recording quality made it difficult to say if the talker from whose speech the stimuli were created used breathy voicing in this word. However, as all steps of the continuum used in the present study were resynthesized from his pronunciation of this word, the initial consonant was the same across the entire continuum, meaning that it should not have played any role in participants' perception of the pairs as the same or different.

	unmodified – unmodified	same
	unmodified – F2 = 1400 Hz	different
	F2 = 1500 Hz – unmodified	different
	unmodified – F2 = 1600 Hz	different
[put] (‘speck’)	unmodified – unmodified	same
	unmodified – unmodified	same
	F2 = 1000 – unmodified	different
	unmodified – F2 = 1100	different
	F2 = 1200 – unmodified	different
	unmodified – F2 = 1300	different
	F2 = 1400 – unmodified	different
	unmodified – F2 = 1500	different

4.3.3.3 Structure of test portion

The structure of the test portion was broken down as follows, and is also summarized below in **Table 4.9**. For the [a]/[ɒ]$_A$ vowel pair, the consistent intervals trials were made up of 2-step intervals, such that step 1 was compared with step 3, step 2 was compared with step 4, etc. For the [ɔ]/[ɒ]$_O$ pair, the consistent intervals trials were made up of 3-step intervals, such that step 1 was compared with step 4, step 2 was compared with step 5, etc. The suitability of these intervals was determined via piloting. 2-step intervals were initially tried for both vowel pairs. For [a]/[ɒ]$_A$, pilot participants performed neither at ceiling nor at floor, so the 2-step intervals were determined to present an appropriate level of difficulty of discrimination. For [ɔ]/[ɒ]$_O$, some pilot participants performed at or near floor with the 2-step intervals, while 3-step intervals seemed to provide a more appropriate level of difficulty of discrimination. Thus, for the consistent intervals trials, each word representing the [a]/[ɒ]$_A$ pair had 5 possible trials (1-3, 2-4, 3-5, 4-6, 5-7), while each word representing [ɔ]/[ɒ]$_O$ had 4 possible trials (1-4, 2-5, 3-6, 4-7).

For the trials with differences of varying magnitudes, both vowel pairs had the same number of possible trials – namely, they both had 6 possible trials for each word (6-7, 5-7, 4-7, 3-7, 2-7, 1-7).

Each of the consistent-size interval and varying-size interval trials was repeated 4 times, so that it would be possible to gauge the consistency of participants' perceptual responses, and so that isolated errors or lapses in attention would not render the data uninformative. The stimuli pairs that overlapped between the two subsets of the trials (5-7 for [a]/[ɒ]$_A$ and 4-7 for [ɔ]/[ɒ]$_O$) were also repeated only 4 times (their inclusion in both subsets of the trials did not mean that they received additional repetitions). The order of the two stimuli within each pair was balanced across these four repetitions – for example, the pair 1-3 was presented twice with the order 1-3 and twice with the order 3-1. As the order in which listeners hear different vowels has been found to affect their discriminability (e.g., Masapollo et al. 2017), it was necessary to account for any such effect by balancing the order of the presentation of the stimuli within each pair.

Because 2-step intervals were tested in the case of the [a]/[ɒ]$_A$ pair, while 3-step intervals were tested in the case of the [ɔ]/[ɒ]$_O$ pair, the [ɔ]/[ɒ]$_O$ pair would have had 4 fewer trials per lexical item given the task structure described above. However, it was preferable for every lexical item to have the same total number of trials, so that the analyses of the different lexical items would be more comparable. For this reason, 4 additional "different" trials for each lexical item representing the [ɔ]/[ɒ]$_O$ pair were added. These trials consisted of the endpoints of each continuum (steps 1 and 7), with the order of the stimuli within the pair balanced across the four trials for each word.

As the [a]/[ɒ]$_A$ and [ɔ]/[ɒ]$_O$ vowel pairs were each represented by two different lexical items in the AX task, the total number of "different" trials (including repetitions) was 80 for each pair. "Same" trials for each lexical item were also presented, in which each step of each continuum was presented together with itself (1-1, 2-2, 3-3, 4-4, 5-5, 6-6, 7-7). Each "same" trial was repeated 4 times, meaning that there were 28 "same" trials per word, or 56 total "same" trials each for [a]/[ɒ]$_A$ and [ɔ]/[ɒ]$_O$. Though it is generally recommended that the number of "different" and "same" trials in an AX task be equal, having more "different" trials is advisable in cases where discrimination is expected to be difficult (McGuire 2010). This is because participants will expect the number of "same" and "different" trials to be equal (McGuire 2010), and we can assume that difficult-to-discriminate "different" trials will often be perceived by

listeners as "same" trials. The strategy of having fewer "same" than "different" trials was determined to be appropriate for the present study precisely for this reason – since the study involves fully or partially allophonic differences that may be very difficult to perceive, especially in cases where participants are presented with small intervals, it could be assumed that having a higher proportion of "different" trials would not violate participants' expectation for equal "same" and "different" trials.

The total number of "different" and "same" trials for the [a]/[ɒ]$_A$ and [ɔ]/[ɒ]$_O$ vowel pairs collectively was thus 272 trials.

Table 4.9: Summary of structure of the test portion of the AX task. See the text for detailed explanations of where each number comes from.

Test trials – 160 "different", 112 "same" (272 total trials)			
[a]/[ɒ]$_A$ (two lexical items; consistent intervals are 2-step intervals)		[ɔ]/[ɒ]$_O$ (two lexical items; consistent intervals are 3-step intervals)	
"Different" trials	"Same" trials	"Different" trials	"Same" trials
80 (40 2-step interval trials; 40 other interval trials)	56	80 (32 3-step interval trials; 48 other interval trials)	56
Filler trials ([a]/[ae]; four lexical items; consistent intervals are 3-step intervals) **– 72 total trials**			
"Different" trials		"Same" trials	
48 (32 3-step interval trials; 16 endpoint pair trials)		24	
Control trials (eight lexical items forming 4 minimal pairs) **– 48 total trials**			
"Different" trials		"Same" trials	
32		16	

Total trials = 392 (240 "different", 152 "same")

Both "same" and "different" trials were also presented for the control condition consisting of the four minimal pairs listed above in **Table 4.7**. In the case of this condition, "different" trials were those in which a minimal pair was presented, while "same" trials were those in which the same word was presented twice. Each "different" trial was repeated 8 times, for a total of 32 "different" trials. The order of the members of each pair were balanced across the trials, such that each word was presented four times as the first member of the pair and four times as the second member of the pair. There were a total of 16 "same" trials (2 repetitions for each word). Thus, the total number of trials for the control condition was 48 trials.

"Same" and "different" trials were also presented from the filler [a]/[ae] vowel pair. As noted above, this vowel pair was represented by 4 different lexical items. The "different" trials for this vowel pair consisted of 3-step intervals (1-4, 2-5, 3-6, 4-7) and endpoint pairs (1-7). For each lexical item, each 3-step interval was repeated twice, with the order of members of the pair balanced across the two repetitions, while the endpoint pair was repeated 4 times, with the order of members of the pair balanced across the four repetitions. Thus, for each lexical item, there were 12 "different" trials. The "same" trials consisted of the pairs 1-1, 4-4, and 7-7, which were repeated twice for each lexical item, resulting in a total of 6 same trials for each lexical item. There were thus a total of 72 trials across the four words representing the [a]/[ae] vowel pair.

4.3.4 Participants and data collection

The AX task was completed by the same 51 participants who completed the production task (see Section 2.4.1 for detailed information about the participants). Information about the data collection procedures for both perception tasks can be found in Section 3.5. As noted in Section 3.5, one participant's responses to both perception tasks were lost.

4.4 Overview of analyses

4.4.1 Criteria for exclusion of participants

Prior to analysis of responses to the AX task, participants' performance on the control condition (obvious minimal pairs) was examined in order to identify any participants who had seemingly not understood the task. Performance on this condition was evaluated using data

visualizations and calculation of the percent of responses that were correct vs. incorrect. One participant (Participant 36) was ultimately excluded from analyses of the AX task due to responding correctly to only around 35% of control trials. The average accuracy on the control trials was 95% (maximum – 100%; minimum not including Participant 36 – 54%). Participant 36 also had a high rate of responding incorrectly to "same" trials of the test stimuli (i.e., responding "different" when the stimuli were actually the same), as examined through visualizations, comparisons of the percent of correct responses across different subsets of stimuli, and *d'* (see Section 4.5.1.2 below). Though Participant 3 (54% correct on control stimuli) can also be considered an outlier in terms of their responses to the control stimuli, their accuracy on the test stimuli did not show the same concerning pattern observed for Participant 36, as their accuracy on "same" trials was consistently high across all the test words. It was thus decided to not exclude Participant 3.

4.4.2 Summary of analyses and predictions

The following sections present the methods and results of two separate sets of analyses. The analyses presented in Section 4.5 involved calculating various individual metrics of performance on the task. Linear regression models were then implemented to explore differences in performance across the two vowel pairs, as well as whether/how demographic factors and metrics related to participants' productions of the relevant vowels predicted performance on the task. The analysis presented in Section 4.6 involved mixed-effects logistic regression models, which investigated how participants' responses to various intervals differed, while also investigating the role of demographic factors and production metrics. In connection with the questions explored by the two subsets of stimuli described above, these models included those predicting performance on intervals of a fixed size at different points along the continua, as well as those predicting responses to intervals representing differences of varying sizes from the canonical SEA variant. In accordance with the core research question explored in this task (i.e., about differences in the discrimination of the different vowel pairs, including differences in discrimination at different points along the continua), all models in both parts of the analysis included predictors such as VOWEL PAIR and STEP PAIR. Some models also included WORD as a predictor (rather than VOWEL PAIR), as the different words representing each vowel pair had different sized acoustic differences in their stimuli, meaning that investigating differences in

their discrimination was also useful for assessing how participants' perceived differences of varying sizes.

As described in Section 4.2 above, it was predicted that the discrimination of differences in the stimuli would differ in some way for the two vowel pairs, in connection with their differing characteristics. However, due to the lack of previous data about perception of the given vowels, the analyses were largely exploratory, and there were various possibilities about how the results for the two vowel pairs might differ. On one hand, the contrastive nature of [ɔ]/[ɒ]$_O$ could allow listeners to discriminate smaller differences than for [a]/[ɒ]$_A$ in the investigation of differences of varying magnitudes. On the other hand, the stimuli for [a]/[ɒ]$_A$ presented larger acoustic differences than those for [ɔ]/[ɒ]$_O$, which could facilitate discrimination of smaller differences (i.e., smaller intervals as measured by the number of continuum steps) for this pair. The social salience of [a]/[ɒ]$_A$ could also play some role in facilitating discrimination of this pair. In the analysis of discrimination of intervals of a consistent size across the entire continuum, it is possible that the contrastive nature of [ɔ]/[ɒ]$_O$ will lead to peaks in perceptual accuracy for this pair (indicating categorical-like perception), while such peaks might be lacking for the allophonic [a]/[ɒ]$_A$ pair. On the other hand, considering that small differences from [a] in the direction of [ɒ]$_A$ may be perceived similarly to a coarticulatory effect (given the findings of Seyfarth and Garellek [2018]), it is possible that participants could have difficulty discriminating this vowel pair near the [a] endpoint (SEA-like endpoint).

Though the analyses thus generally had an exploratory nature, with various possible results that could relate to the differing characteristics of the two vowel pairs, the testing of predictors related to demographic factors and production metrics can be considered the most exploratory aspect of the given analyses, as there were no specific expectations about the potential role of these predictors (in contrast to the VOWEL PAIR, STEP PAIR, and WORD predictors).

In connection with the exploratory nature of the analyses, and also with the lack of a priori reasons for choosing certain analysis methods over others, a variety of possible methods were explored (e.g., using different metrics to evaluate accuracy of discrimination, running pooled vs. separate models for the two vowel pairs), and the results of different possible methods are compared. Though the conventional significance threshold of $p<0.05$ was used, predictors with p-values of less than around 0.2 are also noted and discussed in the sections below (such

predictors are called "sub-significant" throughout this chapter and Chapter 5). Throughout the analysis process, it was noted that decisions about analysis metrics and analysis structure sometimes had substantial effects on which predictors reached vs. only approached p=0.05, such that not taking note of sub-significant trends could have led to a misinterpretation of the data or to an impression that the results of different analyses were more different from one another than they actually were. The threshold of $p<0.2$ (rather than the more widespread marginal significance threshold of 0.1) as an indication that trends should be noted and examined emerged through the analysis itself, as it was observed that ignoring values greater than p=0.1 would have sometimes led to a false impression about the dissimilarity of different models (in cases where, for example, the p-value of a given predictor was slightly above 0.1 in one model and slightly below it in another model). Though some of the identified sub-significant trends may ultimately not be interpretable or meaningful, they are considered indicative of a potentially important tendency in the data in those cases where multiple analyses converge on the same trend (or where sub-significant trends in one analysis correspond to significant ones in another analysis). As such, summaries of the results of the two sets of analyses are presented in Sections 4.5.3.3 and 4.6.2.3, and Section 4.7 further synthesizes these results by identifying the trends that were found numerous times in different analyses.

4.5 Analysis and results – individual metrics of performance and linear regression models

4.5.1 Calculating metrics of individual performance

4.5.1.1 ACCURACY (proportion of correct responses on "different" trials)

Several metrics were calculated in order to assess each participant's performance in the discrimination of the test stimuli. The first of these was the proportion of correct responses among various subsets of the data, which is commonly used as a metric of performance in perception experiments (McGuire 2010: 13), and has been used by previous authors in the analysis of AX tasks involving vowels (e.g., Amengual & Chamorro 2015; Butcher 2022). Following Amengual and Chamorro (2015: 218) and Butcher (2022: 174-175), the proportion of correct responses was calculated only among those trials where the correct answer was "different." This approach assumes that it is participants' responses to "different" trials that should be used to evaluate their performance on the task, rather than their responses to both "different" and "same" trials. Though the aforementioned authors do not explain the reasoning

behind this analytical choice, it can perhaps be assumed to represent a viewpoint according to which the "same" trials are essentially control trials. According to this view, the responses to these trials would not be relevant for evaluating each participant's discrimination, provided that the participant did not show an extremely high rate of errors on the "same" trials indicative of not understanding the task (like Participant 36 mentioned above).

The proportion of correct responses was calculated separately for each word. Two separate metrics were calculated – the proportion of correct responses on all "different" trials for the given word (henceforth referred to as ACCURACY) and the proportion of correct responses only on the subset of trials that presented intervals of a consistent size across the vowel continuum (henceforth referred to as CONSISTENT INTERVALS ACCURACY). Both of these metrics were calculated in order to be able to assess each participant's performance on the consistent intervals trials separately from their general performance on each word. It was desirable to calculate a separate metric for the consistent intervals trials because they represent a separate research question – namely, they aimed to assess participants' discrimination of intervals of a relatively small size at different points along the continuum. The two metrics were also calculated for each of the two vowel pairs ([a]/[ɒ]$_A$ and [ɔ]/[ɒ]$_O$), by pooling the data for the two words for each vowel pair. These pooled metrics are used in some analyses presented in Chapter 5.

4.5.1.2 A measure of accuracy on both "different" and "same" trials – d'

Because the calculations of ACCURACY presented in the previous section do not account for participants' performance on the "same" trials, there is nevertheless a risk that they inflate the performance of participants who may have answered correctly to some "different" trials because of mistakes, random guessing, or an overall higher tendency to choose "different." The metric *d'* offers a solution to this problem by incorporating both accuracy ("hits") and "false positives" (see McGuire 2010: 13), which in the present study correspond to correct answers to "different" trials and incorrect answers (i.e., answering "different") to "same" trials. The metric thus accounts for the possibility that some participants may have achieved a high accuracy on the "different" trials due to a higher overall tendency for choosing "different" even when stimuli sounded identical or close to identical. Though some authors (e.g., McGuire 2010: 13) consider *d'* to be superior to other measures of accuracy, in the context of the present study it was deemed that both raw accuracy on "different" trials and *d'* should be calculated and compared. As the

per-speaker per-word sample size for "same" trials was only 28 trials, *d'* carried the risk of inflating the cost of a small number of mistakes (i.e., of penalizing speakers too harshly for making only a few mistakes on the "same," trials, which would make up a relatively large percentage of the trials due to the small number of observations per participant per word).

The metric *d'* was calculated according to the method described by Macmillan and Creelman (2005: 8) – namely, by calculating the difference between the z-scored proportions of correct "different" trials and incorrect "same" trials. These proportions (i.e., the proportion of correct "different" trials among all the "different" trials, and the proportion of incorrect "same" trials among all the "same" trials) were calculated for each participant for each word and vowel pair (pooled), and were then converted to (by-participant) z-scores in Microsoft Excel using the NORM.S.INV function before being subtracted from one another. Proportions of exactly 0 or 1 were adjusted using the method outlined by Macmillan and Creelman (2005: 8) because it is not mathematically possible to calculate *d'* from such proportions. Namely, proportions of 0 were changed to 1/(2N) and proportions of 1 were changed to 1-1/(2/N) (where N is the number of trials from which the proportion was calculated). This metric was also calculated with pooled data for each vowel pair, as the pooled *d'* values are used in some analyses presented in Chapter 5.

4.5.1.3 SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED

As described in Section 4.2, the goal of the subset of trials that presented intervals of varying sizes was to determine the acoustic distance from the canonical SEA vowels ([a] or [ɔ]) that was needed in order for each listener to start perceiving a difference. Accordingly, a metric of performance on this subset of the trials was extracted from each participant's responses for each word. This metric is henceforth referred to as SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED. Since the goal of this analysis was to determine the interval size at which each participant began to consistently discriminate the stimuli for each word, it was determined that the metric should be defined as the first interval (among the intervals ranging from 6-7 to 1-7) at which the participant reached at least 75% accuracy without dropping to a lower accuracy at any of the larger intervals. For example, if a participant first achieved 75% or higher accuracy at the interval 3-7 and did not drop below 75% accuracy for the intervals 2-7 and 1-7, then their SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED for the given word would be recorded as 3-7. 75% was chosen as the most suitable accuracy threshold because 50% accuracy could be

achieved by chance (and thus would not be an informative threshold), while 100% accuracy would be too harsh of a criterion, not allowing participants to make mistakes. (Intermediate percentages between 50% and 75% or 75% and 100% could not have been chosen because there were only 4 repetitions of most stimuli pairs). **Figures 4.3**-**4.5** illustrate the processes of determining the SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED for each participant for each word. These figures and all other data visualizations in this chapter were created in R (R Core Team 2021) with *ggplot2* (Wickham 2016).

For participants who never reached consistent 75% accuracy at any of the given intervals, the metric was recorded as 1-7+. Each participant's SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED for each word was also recorded as a numeric value representing the size of the interval, so that the metric could be used in some analyses that would require it to be a continuous numeric variable. For example, if a participant reached 75% accuracy at the interval 1-7 for a given word, their interval for that word would also be recorded as 6 (because they discriminated a 6-step interval). For those who never reached 75% accuracy (i.e. those for whom 1-7+ was recorded), the numeric interval was recorded as 7, i.e. a decision was made to code these participants as having required one additional step to reach 75% accuracy in comparison with those who achieved such accuracy at the interval 1-7. Although this decision makes an assumption about how these participants would have performed had the continuum been larger, given that it is not possible to actually know how large of a difference would have been needed in order for them to reach 75% accuracy, it was decided that making such an assumption was preferable to excluding large numbers of participants from various analyses.

Each participant's SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED for each word was also recorded in Bark (i.e., using the stimuli Bark values presented in **Tables 4.4**-**4.5** to calculate the Euclidean distance between the relevant steps), with the goal of performing analyses that would be able to account for the fact that the size of the continua steps was different within and especially across the two vowel pairs. In other words, converting the intervals to Bark would allow for comparison of the actual sizes of the intervals that participants discriminated across the different words and vowel pairs. For participants who never reached 75% accuracy for a given word, the imagined 7-step interval was converted to Bark by simply adding another step (based on the values in **Table 4.6**) to the Bark value for the 6-step interval.

Some analyses presented in Chapter 5 also required pooled metrics for each vowel pair. Accordingly, each participant's numeric intervals and Bark intervals for the two words representing each vowel pair were averaged, and the result was recorded for use in these analyses.

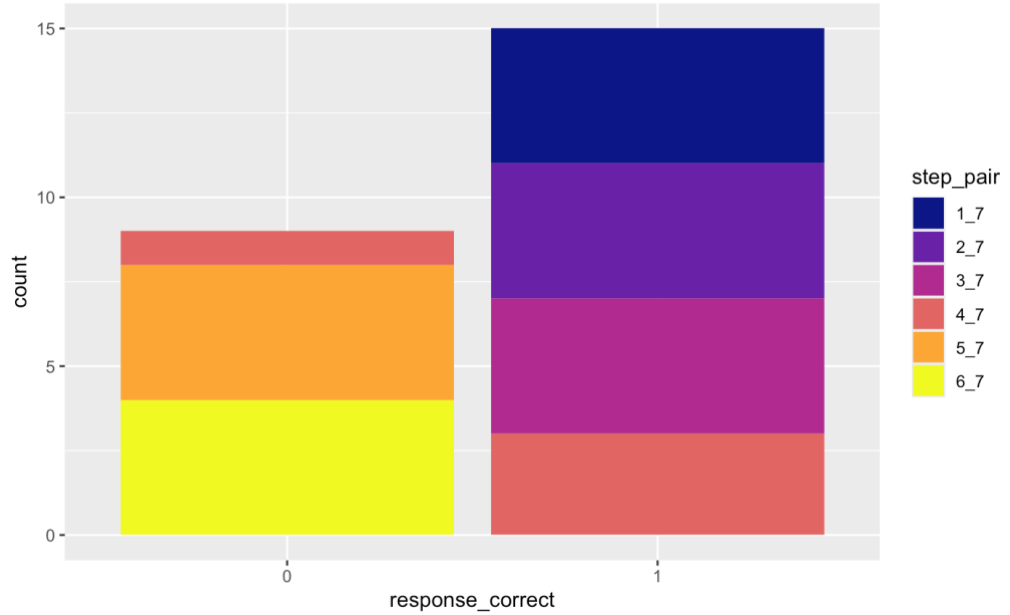


Figure 4.3: Participant 2's responses for varying-size intervals the word *pat*. The left and right-hand columns indicate incorrect and correct responses respectively. The interval 4-7 was chosen because the participant reached 75% accuracy at this interval and had 100% accuracy for all the larger intervals.

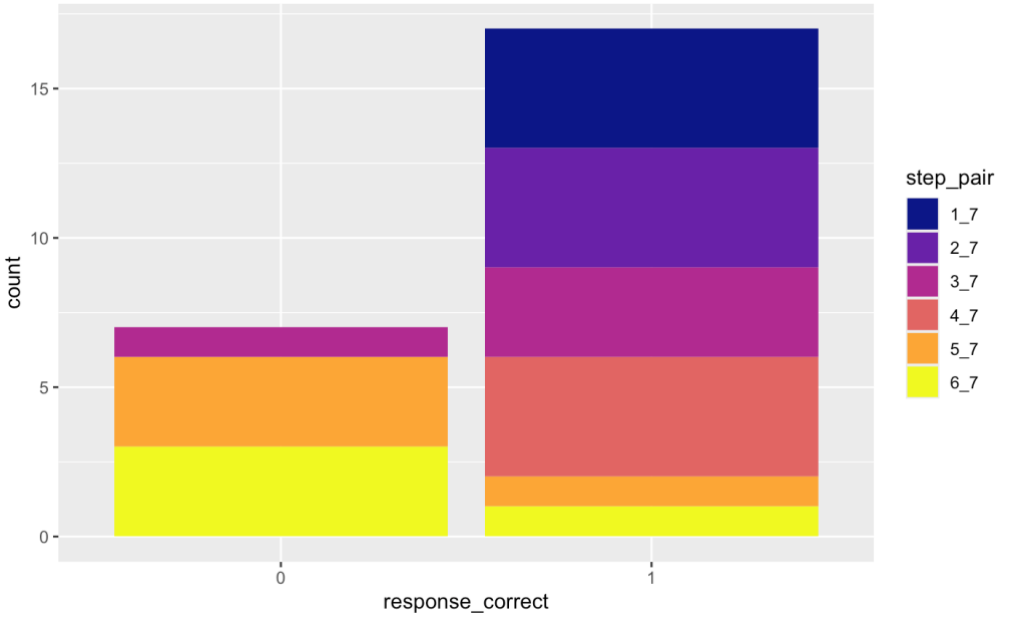


Figure 4.4: Participant 42's responses for varying-size intervals for the word *pat*. The left and right-hand columns indicate incorrect and correct responses respectively. The interval 4-7 was chosen because the participant reached 100% accuracy at that interval (after only having 25% accuracy at the previous interval). Though the participant's accuracy dropped at the subsequent interval 3-7, this did not affect the interval choice because the participant's accuracy at this interval was still 75%.

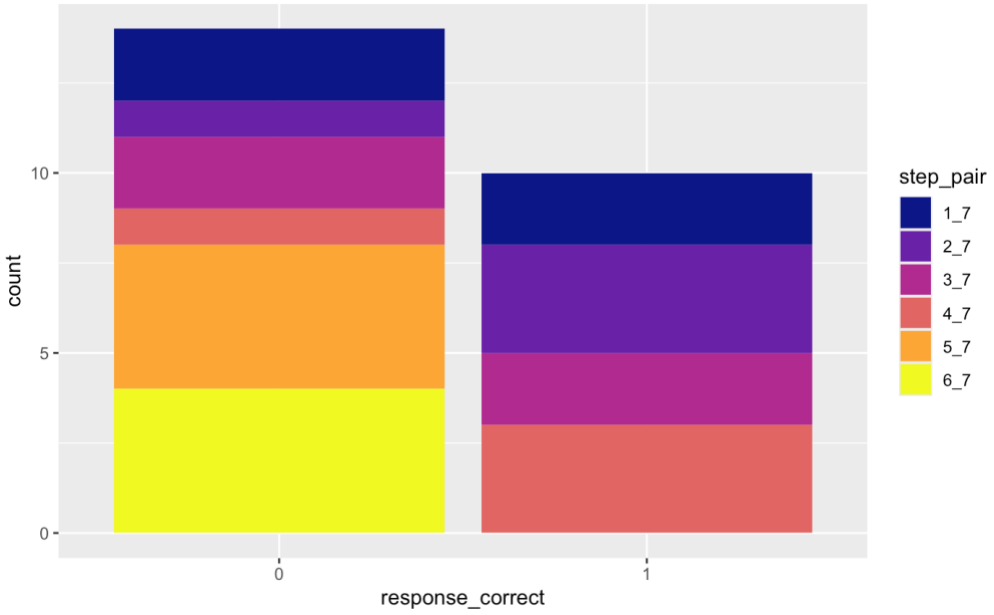


Figure 4.5: Participant 34's responses for varying-size intervals for the word *pat*. The left and right-hand columns indicate incorrect and correct responses respectively. This participant failed to reach consistent 75% accuracy at any interval – although they reached 75% accuracy at the interval 4-7, they dropped to 50% accuracy at the larger intervals 3-7 and 1-7, meaning that their interval was recorded as 1-7+.

4.5.2 Linear regression models – predicting metrics of performance with demographics and production metrics

Mixed-effects linear regression models were used to explore trends found in the AX performance metrics described above. In particular, the goal of running these models was both to explore differences between the different words and vowel pairs identified via visualizations of the performance metrics and to explore whether any demographic factors (GENDER, BIRTH YEAR, EDUCATION LEVEL, PREVIOUS, HATSARAT) or production metrics (Pillai scores and ED-median) predicted overall performance on the AX task. While the logistic regression models described in Section 4.6 below predict correct vs. incorrect responses for different step pairs, the linear regression models described in this section explore the relationships between various predictors and more global metrics of performance on the AX task. Since they do not have STEP PAIR as a predictor, these simpler models have the advantage of being able to directly compare the two

vowel pairs within the same model, which was not advisable with the models presented in Section 4.6 due to the fact that the actual size of the differences presented in different step pairs was very different across the two vowel pairs.

All models were run in R using the *lmer()* function from *lme4* package (Bates et al. 2015), while p-values (Satterthwaite's method) were computed using the *lmerTest* package (Kuznetsova et al. 2017). Models were run separately for five different outcome variables: ACCURACY, CONSISTENT INTERVALS ACCURACY, *d'*, SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED (numeric), and SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED in Bark. All models included random intercepts for participants, which were intended to account for individual participants' different baseline performance on the various outcome metrics. As the data sets used in these models are relatively small, the inclusion of random slopes led to singularity issues, and thus random slopes were not included in any models. Including each participant's Pillai scores and ED-median values as predictors was an additional way that individual variation was explored in these analyses. Continuous predictors were z-scored (centered and scaled) using the *scale()* function in R, as the different continuous predictors being tested (e.g., BIRTH YEAR and Pillai scores) had very different original scales. Collinearity of predictors for each model was tested using the *vif()* (variance inflation factors) function from the *car* package (Fox & Weisberg 2019).

For each outcome variable, models were run with the data for the two vowel pairs both pooled and separated. The only outcome variable for which a pooled model was not run was SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED (numeric), as these intervals are not directly comparable across the two vowel pairs due to large differences in the actual acoustic differences presented in the stimuli for the two vowel pairs (see **Table 4.6**). Models with pooled data were run either with WORD as a predictor (i.e., testing differences between each of the four stimuli words) or with VOWEL PAIR as a predictor (i.e., testing differences between the two vowel pairs overall). Both predictors were tested, but in separate models, as perfect collinearity prohibited them from being included in the same model. In pooled models with the VOWEL PAIR predictor, interactions of VOWEL PAIR and demographic and production predictors were tested, in order to explore whether relationships between these predictors and the outcome variable differed across the two vowel pairs. On the other hand, models with separated data included only the WORD predictor and did not test interactions. For pooled models with the WORD predictor, the

emmeans() and *pairs()* functions from the *emmeans* package (Lenth & Piaskowski 2026) were used to compute pairwise comparisons of all four words with Tukey-adjusted p-values. *Emmeans()* was also used to test the main effect of VOWEL PAIR in models that included dummy-coded categorical predictors interacting with VOWEL PAIR.

The decision to run both pooled and separate models was based primarily on the risk of collinearity of the production metric predictors (Pillai score and ED-median) with the WORD or VOWEL PAIR predictors in the pooled models. In other words, because the range of Pillai scores and ED-median values was generally different for the two vowel pairs, there was a risk that these predictors would be too correlated with one another to be included in the same model. Inclusion of both Pillai scores and ED-median values in any given pooled model indeed resulted in serious collinearity issues, as these predictors are correlated not only with WORD / VOWEL PAIR, but also with one another. Each production predictor was thus tested one at a time in pooled models, which resulted in more reasonable but still potentially concerning GVIF values (i.e., greater than 5, but less than 10 – see Nahhas [2025, section 5.21] for an explanation of how to interpret the *vif()* output). Due to these collinearity issues, it was determined that the separate models should nevertheless also be run. In these separate models, Pillai scores and ED-median could be tested as predictors in the same model, as *vif()* did not indicate serious issues with collinearity (VIF values were always less than 5). Production metrics calculated with both Bark-converted and non-Bark-converted data were tested, though this did not make a difference in the results of any model. Results of both pooled and separate models are presented in Section 4.5.3.2.

As discussed above, the WORD or VOWEL PAIR predictors were included in all models even when they were not significant, as the core goal of running these models was to explore the statistical significance of patterns observed across the different words and vowel pairs in visualizations of the production metrics. On the other hand, because there were no specific expectations regarding relationships between production metrics and demographic factors with performance on the AX task, these predictors were removed from models in cases where they were clearly not significant. In cases where predictors approached significance (i.e., having a p-value lower than around 0.2), these predictors were kept in the models, and models with the given predictors will be presented and discussed in Section 4.5.3.2 below (such predictors are listed as "sub-significant"). This approach of presenting the results was chosen because of the exploratory nature of this portion of the analysis. Unlike the analysis of the production data

presented in Chapter 2, which used AICc and BIC to identify the model(s) with the best balance of fit and parsimony, this analysis of AX performance had the goal of exploring the data in different possible ways (e.g., though both pooled and separate models, and through different metrics of evaluating performance on the task) and identifying potential trends. Trends that were sub-significant in one model sometimes ended up being significant in other models (e.g., in the pooled but not the separate model, or vice versa), and so it was determined that such trends should be noted and discussed in order to compare the outcomes of different (i.e., pooled and separate) models in the most thorough and transparent way possible.

In summary, the mixed-effects linear regression models predicting the five variables listed above, which were run with the data for the two vowel pairs both pooled and separated, had the following structure:

- Random intercepts for participants
- VOWEL PAIR or WORD as a fixed predictor
- (Optionally) Demographic factors as fixed predictors (GENDER, BIRTH YEAR, EDUCATION LEVEL, PREVIOUS, HATSARAT)
- (Optionally) Production metrics as fixed predictors (Pillai scores and ED-median)

4.5.3 Results of analysis of individual metrics of performance

4.5.3.1 Visualizations

This section presents visualizations of the performance metrics described in Section 4.5.1. All visualizations were created with *ggplot2*.

4.5.3.1.1 ACCURACY: proportion of correct responses among "different" trials

Figures 4.6-4.7 visualize ACCURACY (**Figure 4.6**) and CONSISTENT INTERVALS ACCURACY (**Figure 4.7**). These graphs show that both metrics were similar across the two vowel pairs. The member of each pair whose stimuli had smaller acoustic differences (*pat* and *tsots*) also had lower accuracy. However, the discrepancy in the size of differences in the stimuli across the two vowel pairs (i.e., the fact that differences presented for [a]/[ɒ]$_{A}$ were larger than those for [ɔ]/[ɒ]$_{O}$) does not appear to have led to a difference in accuracy across the two vowel pairs.

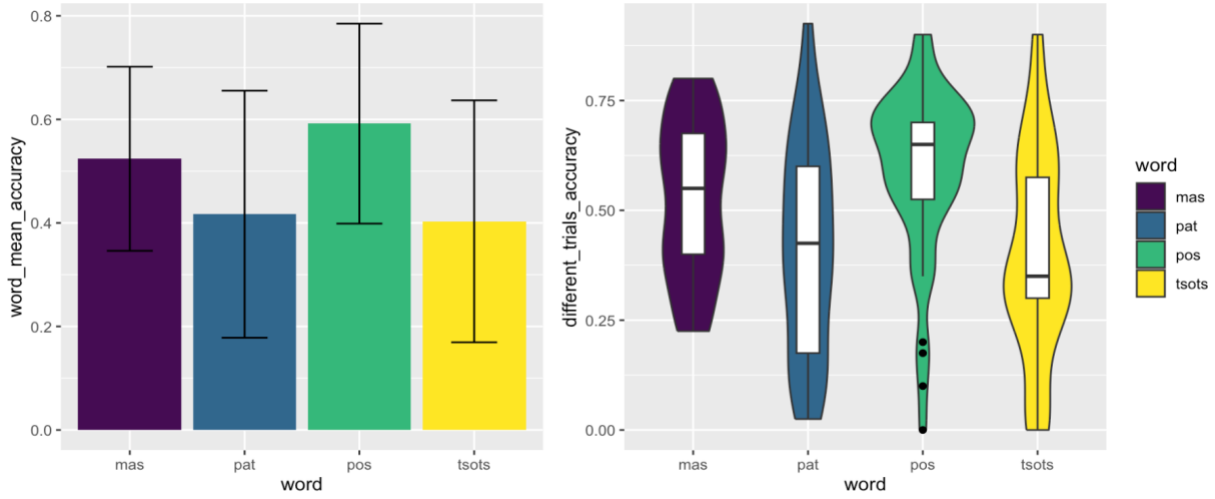


Figure 4.6: The left panel shows mean ACCURACY for all participants by WORD. Error bars represent ± one standard deviation. The right panel shows violin plots of participants' ACCURACY by word with overlaid boxplots. The box represents the range between the first and third quartiles (interquartile range), while the horizontal line represents the median. The whiskers represent data points that lie up to 1.5 times the interquartile range from the edge of the box, while black dots represent outliers.

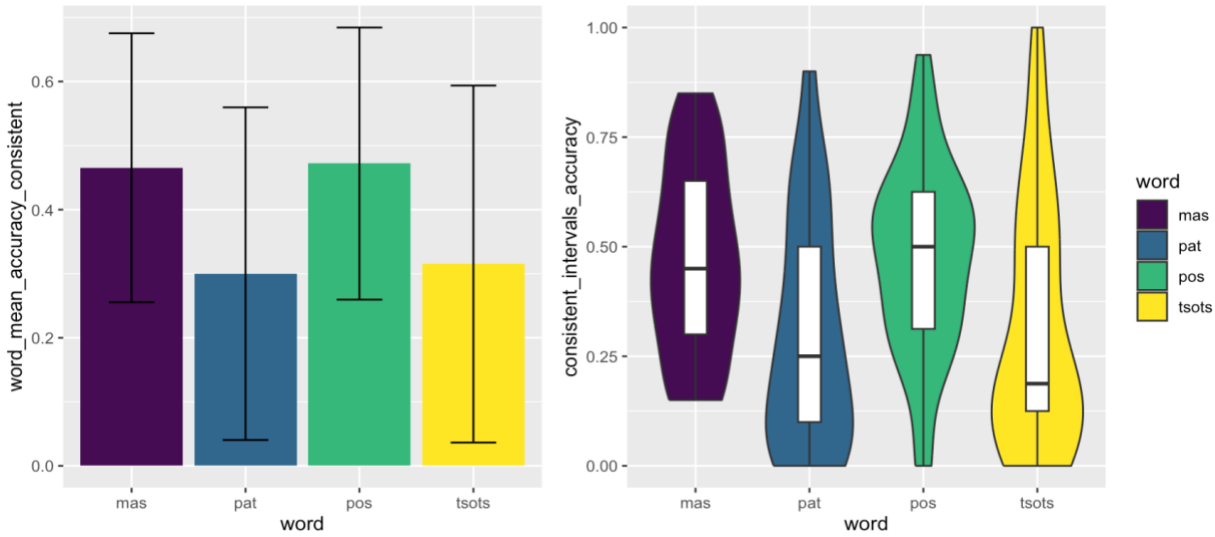


Figure 4.7: The left panel shows mean CONSISTENT INTERVALS ACCURACY by WORD. Error bars represent ± one standard deviation. The right panel shows violin plots of participants' CONSISTENT INTERVALS ACCURACY by word with overlaid boxplots. The box represents the range between the first and third quartiles (interquartile range), while the horizontal line represents the median. The whiskers represent data points that lie up to 1.5 times the interquartile range from the edge of the box.

4.5.3.1.2 d'

Figure 4.8 below visualizes participants' *d'* values for each word. These graphs are similar to **Figures 4.6-4.7** (which present ACCURACY and CONSISTENT INTERVALS ACCURACY), in

that *d'* is similar across the two vowel pairs, but lower for the member of each vowel pair that had smaller acoustic differences in its stimuli (*pat* and *tsots*).

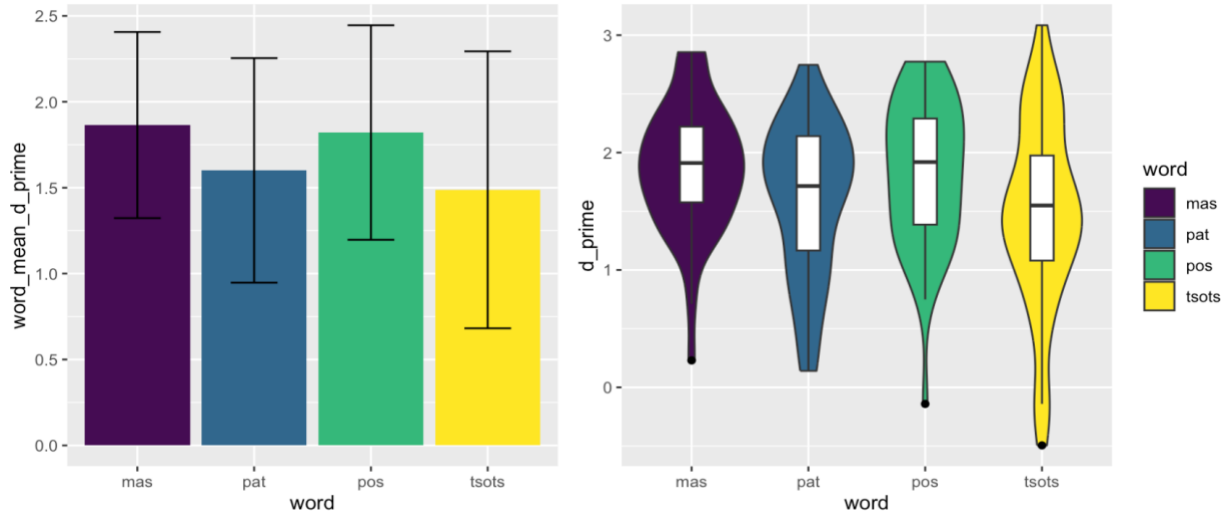


Figure 4.8: The left panel shows mean *d'* values among participants by WORD. Error bars represent ± one standard deviation. The right panel shows violin plots of participants' *d'* values per word with overlaid boxplots. The box represents the range between the first and third quartiles (interquartile range), while the horizontal line represents the median. The whiskers represent data points that lie up to 1.5 times the interquartile range from the edge of the box, while black dots represent outliers.

4.5.3.1.3 SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED

Figure 4.9 presents each participant's SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED, while **Figure 4.10** presents these intervals converted to Bark using the procedure described in Section 4.5.1.3. In both graphs, a higher value corresponds to worse discrimination (i.e., the SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED was larger).

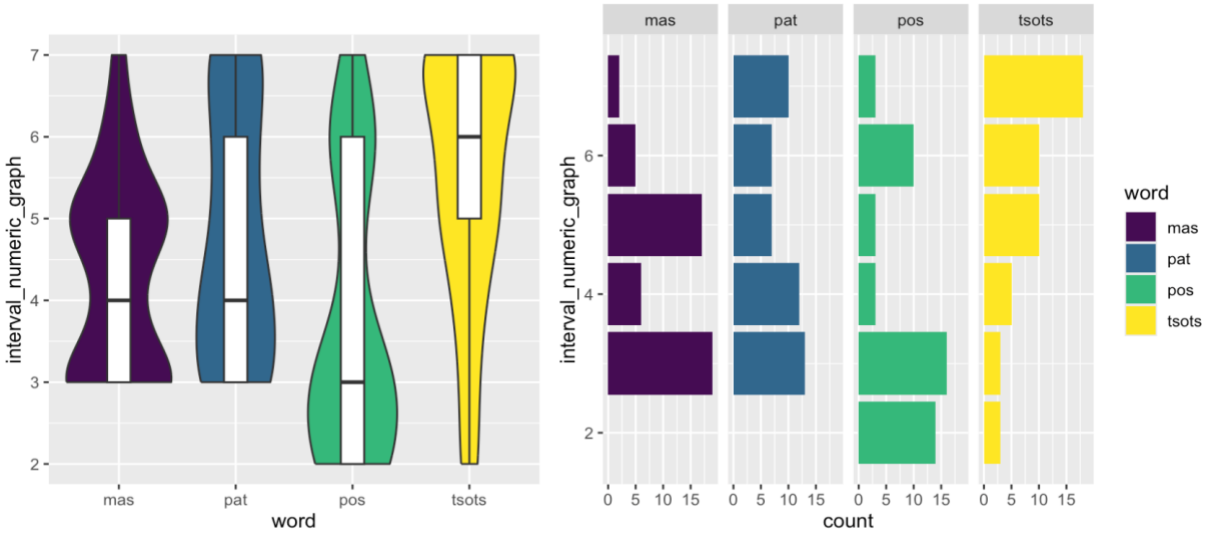


Figure 4.9: The left panel shows violin plots with overlaid boxplots showing each participant's SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED by WORD. A y-axis value of 7 corresponds to an interval of 1-7+ (i.e., not reaching 75% accuracy at any interval), while a y-axis value of 2 corresponds to the interval 5-7 (no participant achieved consistent 75% accuracy at the interval 6-7 for any word). The box represents the range between the first and third quartiles (interquartile range), while the horizontal line represents the median. The whiskers represent data points that lie up to 1.5 times the interquartile range from the edge of the box. The right panel shows the same data as discrete counts of participants assigned to each numeric interval.

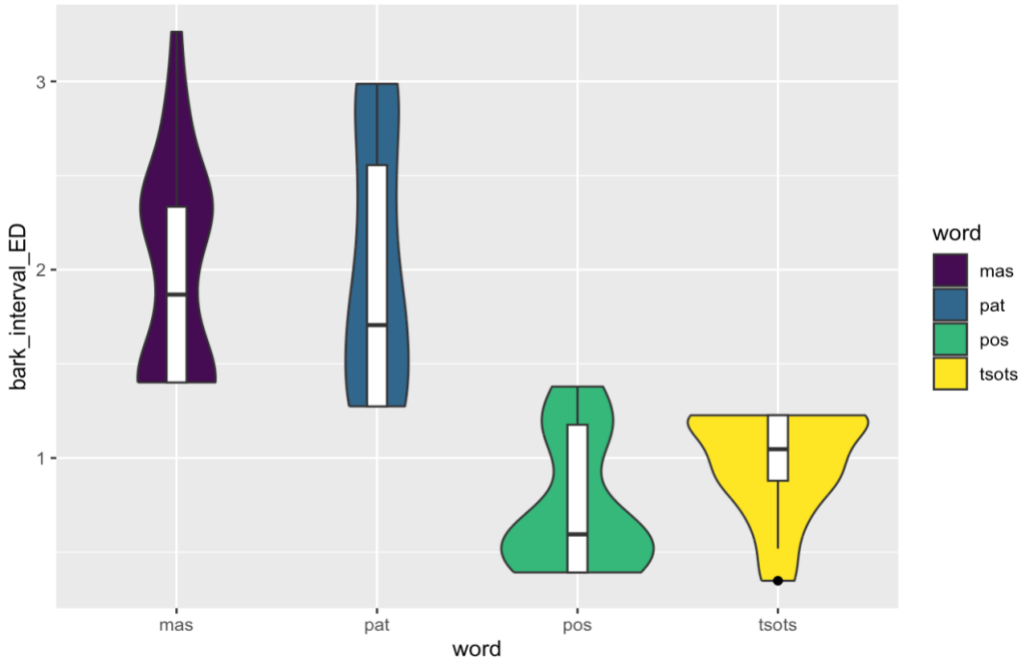


Figure 4.10: Violin plots with overlaid boxplots showing each participant's SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED by WORD in Bark. The box represents the range between the first and third quartiles (interquartile range), while the horizontal line represents the median. The whiskers represent data points that lie up to 1.5 times the interquartile range from the edge of the box, while black dots represent outliers.

As the graphs show, the range of intervals was similar within but not across the two vowel pairs. Though the distribution of intervals among participants is different for *pos* and *tsots*, with many participants discriminating very small intervals for *pos* and many participants having very poor discrimination for *tsots*, there were participants who consistently discriminated the 2-step interval 5-7 for both words. In contrast, no participants discriminated this interval for *pat* and *mas.* As **Figure 4.10** shows, this pattern contrasts with the fact that the acoustic differences in the stimuli were much smaller for *pos* and *tsots* than for *pat* and *mas.*

4.5.3.2 Results of linear regression models

Production metrics (Pillai scores and ED-median, calculated with either Bark-converted or non-Bark-converted data) were never found to be significant predictors of any of the metrics

explored in this section. The subsections below review whether WORD and VOWEL PAIR were found to be significant in predicting the various AX performance metrics, comparing these model results to the patterns identified in visualizations in the previous section. The following subsections also review demographic predictors that were found to be significant or sub-significant in predicting various metrics.

4.5.3.2.1 Predicting ACCURACY

In the separate models for [a]/[ɒ]$_A$, WORD was found to be a significant predictor, such that *pat* had lower ACCURACY than *mas*. EDUCATION LEVEL was also found to be significant, such that people without higher education had higher ACCURACY than those with higher education. A sub-significant effect was found for BIRTH YEAR, such that younger people were less accurate than older people. The model with these predictors is summarized in **Table 4.10**.

Table 4.10: Results of mixed-effects linear regression model predicting ACCURACY for [a]/[ɒ]$_A$. The model included random intercepts for participants. The model structure was as follows: **ACCURACY ~ WORD + EDUCATION LEVEL + BIRTH YEAR + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places.

Predictor	**p-value**
WORD (*pat* vs. *mas)*	*< 0.001*
EDUCATION LEVEL – non-higher vs. higher	*0.030*
BIRTH YEAR	0.095

In the separate models for [ɔ]/[ɒ]$_O$, WORD was also a significant predictor, such that *tsots* had lower ACCURACY than *pos*. HATSARAT was sub-significant, such that people from Hats^{h}aṛat had higher ACCURACY. The model with these predictors is summarized in **Table 4.11**.

Table 4.11: Results of mixed-effects linear regression model predicting ACCURACY for [ɔ]/[ɒ]$_O$. The model included random intercepts for participants. The model structure was as follows: **ACCURACY ~ WORD + HATSARAT + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places.

Pair	**p-value**
WORD (*tsots* vs. *pos)*	*< 0.001*
HATSARAT	0.097

In pooled models with WORD, significant differences were found between the words, while EDUCATION LEVEL and BIRTH YEAR were found to be sub-significant (with people without higher education having higher ACCURACY and younger people having lower ACCURACY). The model with these predictors is summarized in **Table 4.12**.[25] Pairwise comparisons of the levels of WORD in this model are presented in **Table 4.13**. These comparisons show that there are significant differences between all pairs except for *mas – pos* and *pat – tsots*. In other words, the results of the comparisons are consistent with the visualizations presented in **Figures 4.6-4.7** (Section 4.5.3.1.1), which show that ACCURACY was similar across the vowel pairs, but different within each vowel pair in accordance with which member of the pair had larger vs. smaller acoustic differences in the stimuli.

Table 4.12: Results of mixed-effects linear regression model predicting ACCURACY for pooled data with WORD predictor. WORD was dummy coded, with *mas* as the reference level. The model included random intercepts for participants. The structure of the model was as follows: **ACCURACY ~ WORD + EDUCATION LEVEL + BIRTH YEAR + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. This table presents p-values from the *lmer()* output, while Table 4.13 below presents Tukey-adjusted p-values from *emmeans()* pairwise comparisons for WORD.

Predictor	**p-value**
WORD (*pat* vs. *mas)*	*< 0.001*
WORD (*pos* vs. *mas)*	*0.014*
WORD (*tsots* vs. *mas)*	*< 0.001*
EDUCATION LEVEL – non-higher vs. higher	0.107
BIRTH YEAR	0.060

Table 4.13: Results of pairwise comparisons of the levels of WORD from the model presented in **Table 4.12**, computed using *emmeans()*. P-values have been adjusted using the Tukey method. P-values of less than 0.05 are italicized.

Predictor	**p-value**
mas – pat	*< 0.001*

[25] The reference level of WORD in all models was chosen arbitrarily, as pairwise comparisons (implemented via *emmeans()*) would nevertheless be necessary to compare all words with all other words regardless of the coding scheme chosen for the models.

mas – pos	0.066
mas – tsots	*< 0.001*
pat – pos	*< 0.001*
pat – tsots	0.958
pos – tsots	*< 0.001*

In pooled models with VOWEL PAIR, this predictor was not found to be significant (see **Table 4.14**), meaning that there was no overall difference in ACCURACY between the vowel pairs. The model presented in **Table 4.15** tested interactions between VOWEL PAIR and all the demographic predictors that were found to be significant or sub-significant in non-pooled models (EDUCATION LEVEL, BIRTH YEAR, HATSARAT). None of these interactions reached significance, though the interaction with EDUCATION LEVEL approached significance, indicating a difference in the effect of education across the two vowel pairs (which is consistent with the results of the separate models, in which this predictor was relevant for [a]/[ɒ]$_A$).

Table 4.14: Results of mixed-effects linear regression model predicting ACCURACY for pooled data with VOWEL PAIR predictor. The model included random intercepts for participants. The structure of the model was as follows: **ACCURACY ~ VOWEL PAIR + EDUCATION LEVEL + BIRTH YEAR + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places.

Predictor	**p-value**
VOWEL PAIR ([ɔ]/[ɒ]$_O$ vs. [a]/[ɒ]$_A$)	0.241
EDUCATION LEVEL – non-higher vs. higher	0.107
BIRTH YEAR	0.060

Table 4.15: Results of mixed-effects linear regression model predicting ACCURACY for pooled data with VOWEL PAIR predictor and interactions between VOWEL PAIR, EDUCATION LEVEL, BIRTH YEAR, and HATSARAT. The model included random intercepts for participants. The structure of the model was as follows: **ACCURACY ~ VOWEL PAIR + EDUCATION LEVEL + BIRTH YEAR + HATSARAT + VOWEL PAIR:EDUCATION LEVEL + VOWEL PAIR:BIRTH YEAR + VOWEL PAIR:HATSARAT + (1 | participant)**. Categorical predictors were dummy coded, such that the coefficients for predictors involved in interactions do not represent main effects. The main effect of VOWEL PAIR was tested using *emmeans()*, and the difference between the vowel pairs was not significant (p=0.121). P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places.

Predictor	**p-value**

VOWEL PAIR ([ɔ]/[ɒ]$_O$ vs. [a]/[ɒ]$_A$)	0.099
EDUCATION LEVEL – non-higher vs. higher	*0.040*
BIRTH YEAR	0.125
HATSARAT	0.876
VOWEL PAIR – EDUCATION LEVEL interaction	0.093
VOWEL PAIR – BIRTH YEAR interaction	0.990
VOWEL PAIR – HATSARAT interaction	0.218

4.5.3.2.2 Predicting CONSISTENT INTERVALS ACCURACY

In the separate models for [a]/[ɒ]$_A$, WORD was found to be a significant predictor, such that *pat* had lower CONSISTENT INTERVALS ACCURACY than *mas*. EDUCATION LEVEL was also found to be significant, such that people without higher education had higher CONSISTENT INTERVALS ACCURACY than those with higher education. The model with these predictors is summarized in **Table 4.16**.

Table 4.16: Results of mixed-effects linear regression model predicting CONSISTENT INTERVALS ACCURACY for [a]/[ɒ]$_A$. The model included random intercepts for participants. The structure of the model was as follows: **CONSISTENT INTERVALS ACCURACY ~ WORD + EDUCATION LEVEL + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places.

Predictor	**p-value**
WORD (*pat* vs. *mas)*	*< 0.001*
EDUCATION LEVEL – non-higher vs. higher	*0.039*

In the separate models for [ɔ]/[ɒ]$_O$, WORD was also found to be a significant predictor, such that *tsots* had lower CONSISTENT INTERVALS ACCURACY than *pos*. No demographic variables were found to be significant predictors for this vowel pair, but three variables (EDUCATION LEVEL, BIRTH YEAR, and HATSARAT) approached significance in some models, with people without higher education having higher accuracy, younger people having lower accuracy, and people from Hats^{h}aṛat having higher accuracy. **Table 4.17** presents a model that includes all three predictors (note that while the p-values for EDUCATION LEVEL and HATSARAT in this model appear quite far from significance, they were lower in some other models that had different

combinations of predictors, e.g., in models that included only one or two of these predictors rather than all three).

Table 4.17: Results of mixed-effects linear regression model predicting CONSISTENT INTERVALS ACCURACY for [ɔ]/[ɒ]ₒ. The model included random intercepts for participants. The structure of the model was as follows: **CONSISTENT INTERVALS ACCURACY ~ WORD + EDUCATION LEVEL + HATSARAT + BIRTH YEAR + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places.

Predictor	**p-value**
WORD (*tsots* vs. *pos)*	*< 0.001*
EDUCATION LEVEL – non-higher vs. higher	0.175
HATSARAT	0.220
BIRTH YEAR	0.076

In pooled models with WORD, significant differences were found between the words, while EDUCATION LEVEL and BIRTH YEAR were found to be sub-significant (with people without higher education having higher CONSISTENT INTERVALS ACCURACY and younger people having lower CONSISTENT INTERVALS ACCURACY). The model with these predictors is summarized in **Table 4.18**. Pairwise comparisons of the levels of WORD in this model are presented in **Table 4.19**. These comparisons show that there are significant differences between all pairs except for *mas – pos* and *pat – tsots*, once again being significant with the visualizations presented in Section 4.5.3.1.1 (**Figures 4.6-4.7**).

Table 4.18: Results of mixed-effects linear regression model predicting CONSISTENT INTERVALS ACCURACY for pooled data with WORD predictor. WORD was dummy coded, with *mas* as the reference level. The model included random intercepts for participants. The structure of the model was as follows: **CONSISTENT INTERVALS ACCURACY ~ WORD + EDUCATION LEVEL + BIRTH YEAR + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. This table presents p-values from the *lmer()* output, while Table 4.19 below presents Tukey-adjusted p-values from *emmeans()* pairwise comparisons for WORD.

Predictor	**p-value**
WORD (*pat* vs. *mas)*	*< 0.001*
WORD (*pos* vs. *mas)*	0.830
WORD (*tsots* vs. *mas)*	*< 0.001*
EDUCATION LEVEL – non-higher vs. higher	0.051

BIRTH YEAR	0.075

Table 4.19: Results of pairwise comparisons of the levels of WORD from the model presented in **Table 4.18**, computed using *emmeans()*. P-values have been adjusted using the Tukey method. P-values of less than 0.05 are italicized.

Pair	**p-value**
mas – pat	*< 0.001*
mas – pos	0.997
mas – tsots	*< 0.001*
pat – pos	*< 0.001*
pat – tsots	0.962
pos – tsots	*< 0.001*

In pooled models with VOWEL PAIR, this predictor was not found to be significant (see **Table 4.20**), meaning that there was no overall difference in CONSISTENT INTERVALS ACCURACY between the vowel pairs. The model presented in **Table 4.21** tested interactions between VOWEL PAIR and all the demographic predictors that were found to be significant or sub-significant in non-pooled models (EDUCATION LEVEL, BIRTH YEAR, HATSARAT). None of the interactions were significant, though the interaction with HATSARAT came somewhat close, indicating that this predictor had a stronger relationship with accuracy for [ɔ]/[ɒ]$_O$ than for [a]/[ɒ]$_A$.

Table 4.20: Results of mixed-effects linear regression model predicting CONSISTENT INTERVALS ACCURACY for pooled data with VOWEL PAIR predictor. The model included random intercepts for participants. The structure of the model was as follows: **CONSISTENT INTERVALS ACCURACY ~ VOWEL PAIR + EDUCATION LEVEL + BIRTH YEAR + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places.

Predictor	**p-value**
VOWEL PAIR ([ɔ]/[ɒ]$_O$ vs. [a]/[ɒ]$_A$)	0.670
EDUCATION LEVEL – non-higher vs. higher	0.051
BIRTH YEAR	0.075

Table 4.21: Results of mixed-effects linear regression model predicting CONSISTENT INTERVALS ACCURACY for pooled data with VOWEL PAIR predictor and interactions between VOWEL PAIR, EDUCATION LEVEL, BIRTH YEAR, and HATSARAT. The model included random intercepts for participants. The structure of the model was as follows: **CONSISTENT INTERVALS ACCURACY ~ VOWEL PAIR + EDUCATION LEVEL + BIRTH YEAR + HATSARAT + VOWEL**

PAIR:EDUCATION LEVEL + VOWEL PAIR:BIRTH YEAR + VOWEL PAIR:HATSARAT + (1 | participant). Categorical predictors were dummy coded, such that the coefficients for predictors involved in interactions do not represent main effects. The main effect of VOWEL PAIR was tested using *emmeans()*, and the difference between the vowel pairs was not found to be significant (p=0.134). P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places.

Predictor	**p-value**
VOWEL PAIR ([ɔ]/[ɒ]$_O$ vs. [a]/[ɒ]$_A$)	0.630
EDUCATION LEVEL – non-higher vs. higher	*0.040*
BIRTH YEAR	0.236
HATSARAT	0.886
VOWEL PAIR – EDUCATION LEVEL interaction	0.369
VOWEL PAIR – BIRTH YEAR interaction	0.485
VOWEL PAIR – HATSARAT interaction	0.103

4.5.3.2.3 Predicting d'

BIRTH YEAR was the only demographic predictor that was significant in any models predicting *d'*, and was significant in both the separate and pooled models, with younger people having lower *d'* values. The separate models for both vowel pairs found significant differences between the two members of each pair (*pat* vs. *mas* and *tsots* vs. *pos*), with those words that had higher accuracy in the previous sections (*mas* and *pos*) also having higher values of *d'*. **Tables 4.22** and **4.23** below summarize the separate models for the two vowel pairs.

Table 4.22: Results of mixed-effects linear regression model predicting *d'* for [a]/[ɒ]$_A$. The model included random intercepts for participants. The structure of the model was as follows: ***d'*** **~ WORD + BIRTH YEAR + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places.

Predictor	**p-value**
WORD (*pat* vs. *mas)*	*0.004*
BIRTH YEAR	*0.048*

Table 4.23: Results of mixed-effects linear regression model predicting *d'* for [ɔ]/[ɒ]$_O$. The model included random intercepts for participants. The structure of the model was as follows: ***d'*** **~ WORD + BIRTH YEAR + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places.

Predictor	**p-value**

WORD (*tsots* vs. *pos)*	*0.008*
BIRTH YEAR	*0.008*

In pooled models with WORD, significant differences were found between some words. The model with these predictors is summarized in **Table 4.24**. Pairwise comparisons of the levels of WORD in this model are presented in **Table 4.25**. These comparisons between words do not correspond with the visualizations in Section 4.5.3.1.2 (**Figure 4.8**) in as straightforward of a manner as the comparisons between words in relation to ACCURACY and CONSISTENT INTERVALS ACCURACY. While *tsots* is still significantly different from *mas* and *pos* according to the pairwise comparisons, *pat* is not (though the p-value for the comparison of *pat* and *mas* is just barely higher than 0.05, and the unadjusted p-value from the *lmer()* output for this comparison is lower than 0.05). Thus, in comparison with the raw measures of accuracy, *d'* seems to have reduced the difference between the results for the word *pat* and those of the words *(mas, pos*) that have the highest accuracy and *d'* values. This could indicate that *mas* and *pos* had relatively high rates of participants choosing "different" when the stimuli were actually the same, which has no effect on calculations of raw accuracy, but which is penalized in the calculation of *d'*.

Table 4.24: Results of mixed-effects linear regression model predicting *d'* for pooled data with WORD predictor. WORD was dummy coded, with *mas* as the reference level. The model included random intercepts for participants. The structure of the model was as follows: ***d'* ~ WORD + BIRTH YEAR + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. This table presents p-values from the *lmer()* output, while Table 4.25 below presents Tukey-adjusted p-values from *emmeans()* pairwise comparisons for WORD.

Predictor	**p-value**
WORD (*pat* vs. *mas)*	*0.010*
WORD (*pos* vs. *mas)*	0.669
WORD (*tsots* vs. *mas)*	*< 0.001*
BIRTH YEAR	*0.009*

Table 4.25: Results of pairwise comparisons of WORD from the model presented in **Table 4.24**, computed using *emmeans()*. P-values have been adjusted using the Tukey method. P-values of less than 0.05 are italicized.

Pair	**p-value**
mas – pat	0.051

mas – pos	0.974
mas – tsots	*0.002*
pat – pos	0.138
pat – tsots	0.683
pos – tsots	*0.007*

In pooled models with VOWEL PAIR, this predictor was not found to be significant (see **Table 4.26**), meaning that there was no overall difference in *d'* between the vowel pairs.

Table 4.26: Results of mixed-effects linear regression model predicting *d'* for pooled data with VOWEL PAIR predictor. The model included random intercepts for participants. The structure of the model was as follows: ***d'* ~ VOWEL PAIR + BIRTH YEAR + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places.

Predictor	**p-value**
VOWEL PAIR ([ɔ]/[ɒ]$_O$ vs. [a]/[ɒ]$_A$)	0.302
BIRTH YEAR	*0.009*

4.5.3.2.4 Predicting SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED

As noted in Section 4.5.2, both pooled and separate models were run for this metric with the intervals converted to Bark, while only separate models were run predicting the numeric intervals. The models for separate vowel pairs predicting numeric intervals had nearly the same results (in terms of significance of predictors) as the separate models predicting intervals in Bark. The only difference was that a significant difference between *pat* and *mas* was found in the models predicting numeric intervals, but not in the models predicting Bark intervals. Thus, only the models predicting Bark intervals will be reported in this section.

As noted above, in the separate models predicting Bark intervals for the [a]/[ɒ]$_A$ vowel pair, WORD was not found to be a significant predictor. BIRTH YEAR was significant (younger people having larger intervals), while EDUCATION LEVEL was sub-significant (people without higher education having smaller intervals). The model with these predictors is summarized in **Table 4.27**.

Table 4.27: Results of mixed-effects linear regression model predicting SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED in Bark for [a]/[ɒ]$_A$. The model included random intercepts for participants. The structure of the model was as follows: **SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED ~ WORD + EDUCATION LEVEL + BIRTH YEAR + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places.

Predictor	**p-value**
WORD (*pat* vs. *mas)*	0.713
EDUCATION LEVEL – non-higher vs. higher	0.076
BIRTH YEAR	*0.017*

WORD was found to be a significant predictor in the separate models for [ɔ]/[ɒ]$_O$, with *tsots* having larger intervals (i.e., worse discrimination) than *pos*. No demographic predictors came close to significance for this vowel pair. This model is presented in **Table 4.28**.

Table 4.28: Results of mixed-effects linear regression model predicting SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED in Bark for [ɔ]/[ɒ]$_O$. The model included random intercepts for participants. The structure of the model was as follows: **SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED ~ WORD + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places.

Predictor	**p-value**
WORD (*tsots* vs. *pos)*	*<0.001*

In pooled models with WORD, significant differences were found between words, while BIRTH YEAR was also significant (younger people having larger intervals), and EDUCATION LEVEL was sub-significant (people without higher education having smaller intervals).
The model with these predictors is summarized in **Table 4.29**. Pairwise comparisons of the levels of WORD in this model are presented in **Table 4.30**. These comparisons showed significant differences between the two vowel pairs, which are consistent with the visualizations in Section 4.5.3.1.3 (**Figures 4.9-4.10**). Within each vowel pair, *pos* and *tsots* were significantly different from each other, but *pat* and *mas* were not. This is also consistent with the visualizations, in which *pat* and *mas* appear to have similar distributions, while *pos* and *tsots* have opposing distributions despite having similar ranges of interval values.

Table 4.29: Results of mixed-effects linear regression model predicting SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED in Bark for pooled data with WORD predictor. WORD was dummy coded, with *mas* as the reference level. The model included random intercepts for participants. The structure of the model was as follows: **SMALLEST**

INTERVAL CONSISTENTLY DISCRIMINATED ~ WORD + BIRTH YEAR + EDUCATION LEVEL + (1 | participant). P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places. This table presents p-values from the *lmer()* output, while Table 4.30 below presents Tukey-adjusted p-values from *emmeans()* pairwise comparisons for WORD.

Predictor	**p-value**
WORD (*pat* vs. *mas)*	0.664
WORD (*pos* vs. *mas)*	*< 0.001*
WORD (*tsots* vs. *mas)*	*< 0.001*
EDUCATION LEVEL – non-higher vs. higher	0.143
BIRTH YEAR	*0.021*

Table 4.30: Results of pairwise comparisons of the levels of WORD from the model presented in **Table 4.29**, computed using *emmeans()*. P-values have been adjusted using the Tukey method. P-values of less than 0.05 are italicized.

Pair	**p-value**
mas – pat	0.972
mas – pos	*< 0.001*
mas – tsots	*< 0.001*
pat – pos	*< 0.001*
pat – tsots	*< 0.001*
pos – tsots	*0.029*

In pooled models with VOWEL PAIR, this predictor was significant (see **Table 4.31**), which is also consistent with the pairwise comparisons of words. The model presented in **Table 4.32** tested interactions between VOWEL PAIR and all the demographic predictors that were found to be significant or sub-significant in non-pooled models (EDUCATION LEVEL and BIRTH YEAR). The interaction with BIRTH YEAR was significant, while the interaction with EDUCATION LEVEL was sub-significant, indicating that the effect of these predictors differed for the two vowel pairs. This is consistent with the separate models, in which these predictors were found to be relevant to the outcome variable for [a]/[ɒ]$_{A}$ but not for [ɔ]/[ɒ]$_{O}$.

Table 4.31: Results of mixed-effects linear regression model predicting SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED in Bark for pooled data with VOWEL PAIR predictor. The model included random intercepts for participants. The structure of the model was as follows: **SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED ~ VOWEL PAIR + BIRTH YEAR + EDUCATION LEVEL + (1 | participant).** P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places.

Predictor	**p-value**
VOWEL PAIR ([ɔ]/[ɒ]$_O$ vs. [a]/[ɒ]$_A$)	*< 0.001*
EDUCATION LEVEL – non-higher vs. higher	0.143
BIRTH YEAR	*0.021*

Table 4.32: Results of mixed-effects linear regression model predicting SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED in Bark for pooled data with VOWEL PAIR predictor and interactions between VOWEL PAIR, EDUCATION, and BIRTH YEAR. The model included random intercepts for participants. The structure of the model was as follows: **SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED ~ VOWEL PAIR + BIRTH YEAR + EDUCATION LEVEL + VOWEL PAIR:BIRTH YEAR + VOWEL PAIR:EDUCATION LEVEL + (1 | participant)**. Categorical predictors were dummy coded, such that the coefficients for predictors involved in interactions do not represent main effects. The main effect of VOWEL PAIR was tested using *emmeans()*, and a significant difference was found between the vowel pairs ($p<0.001$). P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places.

Predictor	**p-value**
VOWEL PAIR ([ɔ]/[ɒ]$_O$ vs. [a]/[ɒ]$_A$)	*< 0.001*
EDUCATION LEVEL – non-higher vs. higher	*0.025*
BIRTH YEAR	*0.002*
VOWEL PAIR – EDUCATION LEVEL interaction	0.052
VOWEL PAIR – BIRTH YEAR interaction	*0.038*

4.5.3.3 Summary of results

Tables 4.33-4.36 below summarize the results of the models presented in Section 4.5.3.2, including information about demographic and production predictors. Interactions between VOWEL PAIR and demographic predictors tested in the pooled models are omitted from these tables, as the results are often redundant with the results of the separate models (see Section 4.5.3.2 for details about results of these interactions).

As the tables show, VOWEL PAIR was not significant in the three sets of models that assessed more global measures of performance (ACCURACY, CONSISTENT INTERVALS ACCURACY, and *d'*), while there was a significant difference between the vowel pairs in the models predicting

SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED (which quantified the size of the difference from the canonical SEA variant [step 7] needed for each participant to reach consistent 75% accuracy in their discrimination). Specifically, participants discriminated smaller intervals for [ɔ]/[ɒ]$_O$ than for [a]/[ɒ]$_A$. As noted in Section 4.3.2.1, the acoustic differences in the stimuli were larger for [a]/[ɒ]$_A$, such that the result of this vowel pair being discriminated similarly or worse than [ɔ]/[ɒ]$_O$ is somewhat unexpected. As will be further discussed in Section 4.7, the worse-than-expected discrimination of [a]/[ɒ]$_A$ can likely be explained by phonological characteristics of the two vowel pairs, as [a]/[ɒ]$_A$ is allophonic, and very small differences from the canonical SEA variant [a] may be perceived as coarticulatory (see Section 3.3). Many of the models (e.g., separate and pooled models for both ACCURACY and CONSISTENT INTERVALS ACCURACY; separate models for *d'*) found significant differences between performance on the two members of each vowel pair, such that discrimination was worse for the word with smaller acoustic differences in the stimuli. Thus, there is evidence that the size of differences presented in the stimuli did affect participants' performance, but this effect seems to have been more pronounced within than across the vowel pairs.

Table 4.33: Summary of results of models predicting ACCURACY.

	[a]/[ɒ]$_A$ separate	**[ɔ]/[ɒ]$_O$ separate**	**pooled – word**	**pooled – vowel pair**
significant	*mas – pat* (*mas* has higher accuracy); EDUCATION LEVEL (less education = higher accuracy)	*pos – tsots* (*pos* has higher accuracy)	*mas – pat*; *mas – tsots*; *pat – pos*; *pos – tsots*	
sub-significant	BIRTH YEAR (older = higher accuracy)	HATSARAT (people from Hats^{h}aṛat = higher accuracy)	EDUCATION LEVEL (less education = higher accuracy); BIRTH YEAR (older = higher accuracy	EDUCATION LEVEL (less education = higher accuracy); BIRTH YEAR (older = higher accuracy)
not significant (among tested word and vowel pair comparisons)			*mas – pos*; *pat – tsots*	VOWEL PAIR

Table 4.34: Summary of results of models predicting CONSISTENT INTERVALS ACCURACY.

	[a]/[ɒ]$_A$ separate	**[ɔ]/[ɒ]$_O$ separate**	**pooled – word**	**pooled – vowel pair**
significant	*mas – pat* (*mas* has higher accuracy); EDUCATION LEVEL (less education = higher accuracy)	*pos – tsots* (*pos* has higher accuracy)	*mas – pat*; *mas – tsots*; *pat – pos*; *pos – tsots*	
sub-significant		HATSARAT (people from Hats^{h}aṛat= higher accuracy); EDUCATION LEVEL (less education = higher accuracy); BIRTH YEAR (older = higher accuracy	EDUCATION LEVEL (less education = higher accuracy); BIRTH YEAR (older = higher accuracy	EDUCATION LEVEL (less education = higher accuracy); BIRTH YEAR (older = higher accuracy)
not significant (among tested word and vowel pair comparisons)			*mas – pos*; *pat – tsots*	VOWEL PAIR

Table 4.35: Summary of results of models predicting *d'*.

	[a]/[ɒ]$_A$ separate	**[ɔ]/[ɒ]$_O$ separate**	**pooled – word**	**pooled – vowel pair**
significant	*mas – pat* (*mas* has higher accuracy); BIRTH YEAR (older = higher accuracy)	*pos – tsots* (*pos* has higher accuracy); BIRTH year (older = higher accuracy)	*mas – tsots*; *pos – tsots*; BIRTH YEAR (older = higher accuracy)	BIRTH YEAR (older = higher accuracy)
sub-significant				
not significant (among tested word and vowel pair comparisons)			*mas – pat*; *mas – pos*; *pat – pos*; *pat – tsots*	VOWEL PAIR

Table 4.36: Summary of results of models predicting SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED in Bark.

	[a]/[ɒ]$_A$ separate	**[ɔ]/[ɒ]$_O$ separate**	**pooled – word**	**pooled – vowel pair**
significant	BIRTH YEAR (older = higher accuracy)	*pos – tsots* (*pos* has smaller intervals)	*mas – pos*; *mas – tsots*; *pat – pos*; *pat – tsots*; *pos – tsots*; BIRTH YEAR (older = higher accuracy)	VOWEL PAIR ([a]/[ɒ]$_A$ has larger intervals); BIRTH YEAR (older = higher accuracy)

sub-significant	EDUCATION LEVEL (less education = higher accuracy)		EDUCATION LEVEL (less education = higher accuracy)	EDUCATION LEVEL (less education = higher accuracy)
not significant (among tested word and vowel pair comparisons)	*mas – pat*		*mas – pat*	

4.6 Analysis and results – logistic regression models

4.6.1 Logistic regression models – predicting responses to various intervals

Mixed-effects logistic regression models were used to evaluate participants' responses to various stimuli pairs in the AX task. Such models have been used in many previous studies involving vowel identification or discrimination (e.g., Fridland & Kendall 2012; Kendall & Fridland 2017; Warburton 2020). The data included in these models consisted only of the "different" trials, and separate models were run for the subset of the data representing intervals of a consistent size and the subset of the data representing intervals of varying sizes. It was necessary to separate the models in this way because the two sets of intervals address different questions about participants' responses. The varying intervals seek to determine the size of the difference from canonical SEA variants that listeners can consistently discriminate, while the consistent intervals seek to determine whether discrimination is easier at some particular point along each continuum.

The models were also separated by vowel pair (i.e., the data for [a]/[ɒ]$_A$ and [ɔ]/[ɒ]$_O$ were included in separate models). Separating the data by vowel pair was necessary for two reasons: 1) For the intervals of a consistent size, the step pairs being compared for the two vowel pairs were different, such that the two vowel pairs could not be included in a single model with the same STEP PAIR predictor, and 2) Since the two vowel pairs occupy different parts of the vowel space and have different sized differences between their continua steps, pooling the data even for the analysis of the varying intervals would not be advisable. Each level of STEP PAIR would mean something very different for the two vowel pairs (e.g., both vowel pairs would include a level "1-7" of this predictor, but for each vowel pair it would represent a different-sized difference and different location in the vowel space). In other words, the STEP PAIR predictor would not have a consistent meaning across the two vowel pairs.

Some previous studies that include perception data where a given vowel continuum is represented by multiple words use separate models for each word (Fridland & Kendall 2012; Kendall & Fridland 2017). For this study, however, it was determined that such an approach would not be suitable, as splitting the models would decrease statistical power, which is already relatively low due to the smaller number of participants and observations per participant. The models were thus split into those with consistent intervals vs. varying intervals data and separated for the two different vowel pairs, but the data for the two words representing a given vowel pair were pooled within a given model. A similar approach was used by D'Onofrio (2021).

The binary dependent variable in all logistic regression models was whether each response was correct (i.e., the participant answered "different," and the stimuli were actually different) or incorrect (i.e., the participant answered "same," but the stimuli were actually different). The predictors were as follows:

- **STEP PAIR –** This was a fixed predictor. In models predicting responses to consistent intervals, its levels were 1-3, 2-4, 3-5, 4-6, and 5-7 for [a]/[ɒ]$_{A}$ and 1-4, 2-5, 3-6, 4-7 for [ɔ]/[ɒ]$_{O}$. For models predicting responses to varying intervals, its levels were 1-7, 2-7, 3-7, 4-7, 5-7, 6-7 for both vowel pairs. These levels were converted to single numeric values (e.g., intervals 1, 2, 3, etc.), so that the STEP PAIR predictor could be implemented as a continuous numeric predictor. The goal of including this predictor was to explore how participants' responses (correct vs. incorrect) differed in response to different stimuli pairs.
- **WORD –** This was a fixed categorical predictor, in which the levels were the two words representing each vowel pair. Model results in relation to this predictor indicate how participants' overall responses (correct vs. incorrect) were different for the different words.
- **Interaction between WORD and STEP PAIR –** This interaction tells us whether the pattern of responses across different step pairs is different for the two different words.
- **Random intercepts for trial order –** This random effect accounts for the influence of the order of the presentation of the trials on the results, by allowing the model to account for the possibility that different trials could have different baseline responses based on their order in the task.

- **Random intercepts for participants –** This random effect accounts for the possibility that different participants could have different baseline levels of responding correctly.
- **Random slopes for participants, relative to STEP PAIR, WORD, and their interaction –** These random slopes account for the fact that the effects of STEP PAIR, WORD, and their interaction on responses could be different for different participants. Random slopes relative to the interaction of STEP PAIR and WORD could not be included in the model predicting responses for consistent intervals for [ɔ]/[ɒ]$_O$, as including these random slopes resulted in singularity issues due to the smaller data set included in this model in comparison with the other three models (consistent intervals for [a]/[ɒ]$_A$, varying intervals for [a]/[ɒ]$_A$, varying intervals for [ɔ]/[ɒ]$_O$).
- **Demographic factors and production metrics –** The same demographic factors and production metrics tested in the linear models described in Section 4.5 were tested in these models as fixed predictors.

The models were implemented in R using the *glmer()* function from the *lme4* package, while p-values were calculated using the *lmerTest* package (Laplace Approximation method). As in the linear models presented above, continuous predictors were scaled. Collinearity was checked using the *vif()* function from the *car* package. Models were checked for convergence issues using the *allFit()* function from the *lme4* package. This function refits the *glmer()* (or *lmer()*) model using all possible optimizers, and if all optimizers result in coefficient values that are nearly equivalent, then convergence warnings from the original *(g)lmer()* output can be ignored (Bates et al. 2015). Different models were compared using AIC and BIC, while models with and without correlations between the random effects in relation to participants were compared using ANOVA. In all cases, correlations between the random effects improved the model fit according to ANOVA and did not lead to convergence issues, so all models presented below included such correlations. As with the linear models described above, demographic factors and production metrics were not kept in the models if they were clearly not significant, while WORD, STEP PAIR, their interaction, and the random effects were included in all models because they represent the core questions that this analysis aimed to test. As with the models presented in Chapter 2, AIC and BIC were used as a general guide for comparing models rather than as a definitive way of choosing a single best model. In cases where these metrics disagreed,

were very close for several different models, or preferred models whose explanatory power seemed dubious, these issues and the relevant models are discussed. As with the linear models presented above, sub-significant trends are also discussed, in connection both with the exploratory nature of the analysis and with recognition of the potential influence of the analysis structure on which trends reach significance. Differences between models that included production metrics calculated with Bark-converted vs. non-Bark-converted data are also discussed when relevant (specifically for cases involving ED-median, as Bark conversion did not influence Pillai values beyond very small effects of rounding, as noted in Chapter 2).

4.6.2 Results of logistic regression models

4.6.2.1 Consistent intervals

4.6.2.1.1 [a]/[ɒ]$_A$

In the mixed-effects logistic regression models predicting responses to consistent intervals for the [a]/[ɒ]$_A$ continuum, no demographic factors or production metrics were found to be significant predictors. Thus, the results of the simplest model (with fixed predictors of STEP PAIR, WORD, and their interaction, along with random intercepts for trial order, random intercepts for participants, and random slopes for participants relative to STEP PAIR, WORD, and their interaction) are presented in **Table 4.37**. As the table shows, STEP PAIR, WORD, and their interaction were all significant in this model. Correct responses were significantly lower overall for *pat* than for *mas*. The STEP PAIR predictor's negative slope indicates that correct responses were lower toward the right side of the continuum (whose endpoint is the interval 5-7) than toward the left side of the continuum (whose endpoint is the interval 1-3) for the reference level *mas*. The significant interaction term indicates that there was a difference in the slope of STEP PAIR for the two words, such that the slope for *pat* was less strongly negative than the slope for *mas*.

Table 4.37: Results of mixed-effects logistic regression model predicting responses to consistent intervals for [a]/[ɒ]$_A$. The model structure was as follows: **RESPONSE CORRECT ~ STEP PAIR * WORD + (STEP PAIR * WORD | participant) + (1 | trial order)**. The reference level of WORD was *mas.* P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places.

Predictor	**estimate**	**p-value**

STEP PAIR	-1.506	*<0.001*
WORD (*pat* vs. *mas)*	-1.158	*0.004*
STEP PAIR – WORD interaction	0.796	*0.042*

The tendencies identified by the model can all be seen in the pooled raw data for these two words, depicted in **Figure 4.11** below, as well as in the more detailed distributions of the data shown in **Figure 4.12**. As **Figure 4.11** shows, there was a sharp drop in correct responses at the interval 5-7 for both words, in comparison with other intervals. (Recall that step 7 represents the canonical SEA variant of the vowel, while step 1 represents the most dialect-like variant).

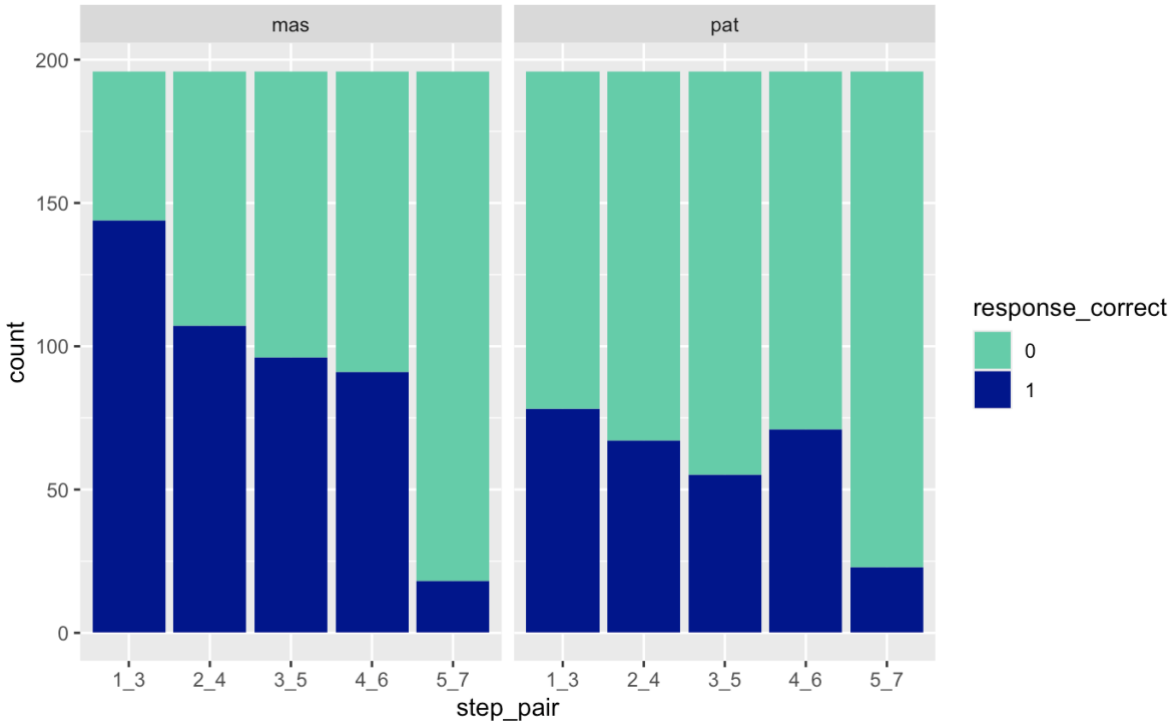


Figure 4.11: Pooled responses by WORD for consistent intervals (2-step intervals) for the [a]/[ɒ]$_A$ vowel continuum. Dark blue represents correct responses (i.e., correctly discriminating the two members of the pair), while green represents incorrect responses (i.e., failing to discriminate the two members of the pair).

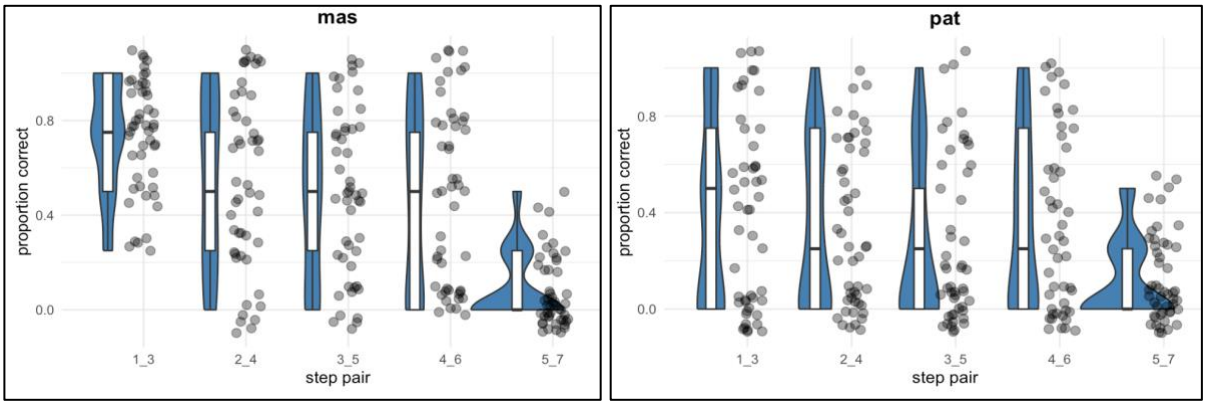


Figure 4.12: Violin plots, boxplots, and jittered scatter plots of individual responses by WORD for consistent intervals (2-step intervals) for the [a]/[ɒ]$_A$ vowel continuum. Each dot represents one participant's proportion of correct responses for the given step pair for the given word (the points have been jittered both vertically and horizontally so that all points would be visible, which results in some points appearing to be below 0 or above 1). The violin plots and boxplots summarize the distribution of the data shown in the scatter plots. The box represents the range between the first and third quartiles (interquartile range), while the horizontal line represents the median. The whiskers represent data points that lie up to 1.5 times the interquartile range from the edge of the box.

4.6.2.1.2 [ɔ]/[ɒ]$_O$

As noted above, the mixed-effects logistic regression models predicting responses to consistent intervals for the [ɔ]/[ɒ]$_O$ continuum could not include random slopes for the interaction between STEP PAIR and WORD, as the inclusion of such random slopes led to issues with singularity. This is likely due to the smaller data set for this vowel pair, which results from the consistent intervals presented in the task consisting of 3-step intervals rather than 2-step intervals. In the models for this vowel pair, no demographic factors or production metrics were found to be significant. However, two demographic factors – EDUCATION LEVEL and BIRTH YEAR – approached significance in some models, with people without higher education and younger people having fewer correct responses. The results of a model including these predictors are presented below in **Table 4.38**. While the two demographic predictors are quite far from significance in this model, they came closer to significance (p = approximately 0.1) in models in which only one of them was included. The model presented below had a very similar AIC (1558.1) to that of the simplest model (with fixed predictors of STEP PAIR, WORD, and their interaction, along with random intercepts for trial order, random intercepts for participants, and random slopes for participants relative to STEP PAIR and WORD), which had an AIC of 1558.2, and the inclusion of the demographic predictors had negligible effects on the significance of the other predictors. As **Table 4.38** shows, STEP PAIR had a positive slope, such that correct responses were higher toward the right side of the continuum (i.e., toward the interval 4-7) for the reference level *pos*. *Tsots* had significantly lower correct responses than *pos* overall. The

coefficient of the interaction term indicates that the slope of STEP PAIR in relation to *tsots* was less steeply positive than the slope in relation to *pos,* though this difference does not quite reach significance (p=0.058).

Table 4.38: Results of mixed-effects logistic regression model predicting responses to consistent intervals for [ɔ]/[ɒ]$_O$. The model structure was as follows: **RESPONSE CORRECT ~ STEP PAIR * WORD + EDUCATION LEVEL + BIRTH YEAR + (STEP PAIR + WORD | participant) + (1 | trial_order)**. The reference level for WORD was *pos.* P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places.

Predictor	**estimate**	**p-value**
STEP PAIR	1.111	*<0.001*
WORD (*tsots* vs. *pos)*	-1.174	*0.008*
EDUCATION LEVEL – non-higher vs. higher	-0.578	0.176
BIRTH YEAR	-0.242	0.190
STEP PAIR – WORD interaction	-0.742	0.058

The trends identified by the model can be seen in the pooled raw data depicted in **Figure 4.13**, as well as in the more detailed distributions depicted in **Figure 4.14**. (Recall that step 7 represents the SEA-like variant of the vowel, while step 1 represents the dialect-like variant).

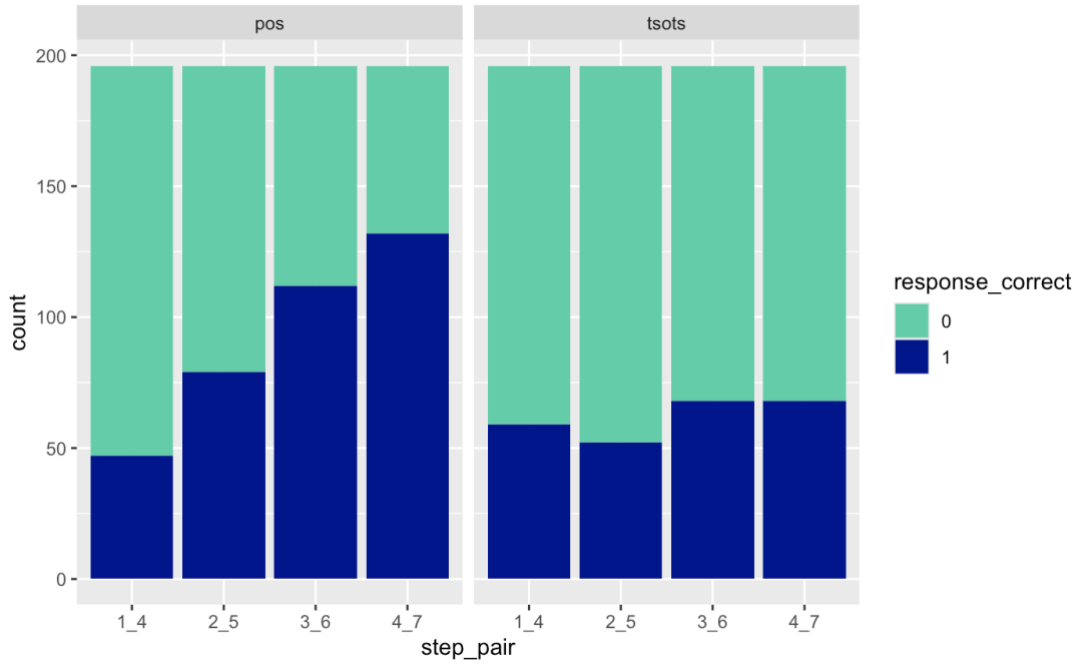


Figure 4.13: Pooled responses by WORD for consistent intervals (3-step intervals) for the [ɔ]/[ɒ]$_O$ vowel continuum. Green represents correct responses (i.e., correctly discriminating the two members of the pair), while red represents incorrect responses (i.e., failing to discriminate the two members of the pair).

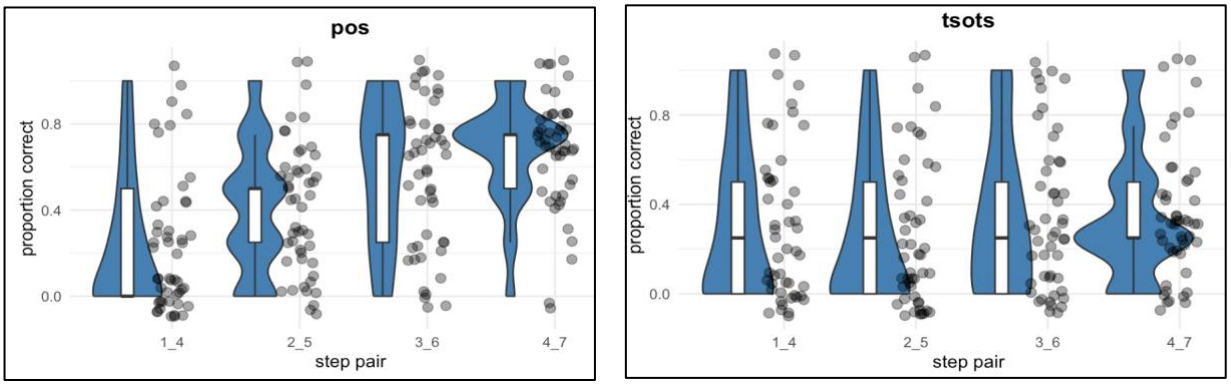


Figure 4.14: Violin plots, boxplots, and jittered scatter plots of individual responses by WORD for consistent intervals (3-step intervals) for the [ɔ]/[ɒ]$_O$ vowel continuum. Each dot represents one participant's proportion of correct responses for the given step pair for the given word (the points have been jittered both vertically and horizontally so that all points would be visible, which results in some points appearing to be below 0 or above 1). The violin plots and boxplots summarize the distribution of the data shown in the scatter plots. The box represents the range between the first and third quartiles (interquartile range), while the horizontal line represents the median. The whiskers represent data points that lie up to 1.5 times the interquartile range from the edge of the box.

4.6.2.2 Varying intervals

4.6.2.2.1 [a]/[ɒ]$_A$

In the mixed-effects logistic regression models predicting responses to varying intervals for the [a]/[ɒ]$_A$ continuum, EDUCATION LEVEL was found to be a significant predictor, with people without higher education having more correct responses. **Table 4.39** shows the results of the model with EDUCATION LEVEL (which also included fixed predictors of STEP PAIR, WORD, and their interaction, along with random intercepts for trial order, random intercepts for participants, and random slopes for participants relative to STEP PAIR, WORD, and their interaction). The slope for STEP PAIR was significant and positive, meaning that correct responses increased as the intervals became larger for the reference level *mas.* WORD was not significant, meaning that the overall level of correct responses was not significantly different across the two words. This lack of a significant difference between the words corresponds with the results of the models predicting SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED in Bark for this vowel pair, presented in Section 4.5.3.2.4, and is also consistent with the visualizations of these intervals presented in Section 4.5.1.3, in which *mas* and *pat* appear to have similar distributions. The significant interaction term indicates that the slope of STEP PAIR was less steeply positive for *pat* than for *mas*.

Table 4.39: Results of mixed-effects logistic regression model predicting responses to varying intervals for [a]/[ɒ]$_A$. The model structure was as follows: **RESPONSE CORRECT ~ STEP PAIR * WORD + EDUCATION LEVEL + (STEP PAIR * WORD | participant) + (1 | trial_order)**. The reference level for WORD was *mas*. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places.

Predictor	**estimate**	**p-value**
STEP PAIR	3.396	*<0.001*
WORD (*pat* vs. *mas)*	0.048	0.905
EDUCATION LEVEL – non-higher vs. higher	0.929	*0.044*
STEP PAIR – WORD interaction	-0.910	*0.031*

The significant interaction term likely reflects the slightly noisier responses for *pat*, which can be seen in **Figures 4.15-4.16**. As the figures show, *pat* had more incorrect responses for the largest intervals and slightly more correct responses for the smallest intervals, making the difference in responses between the largest and smallest intervals less dramatic. **Figure 4.17** depicts pooled responses by EDUCATION LEVEL, showing a higher proportion of correct responses among those without higher education.

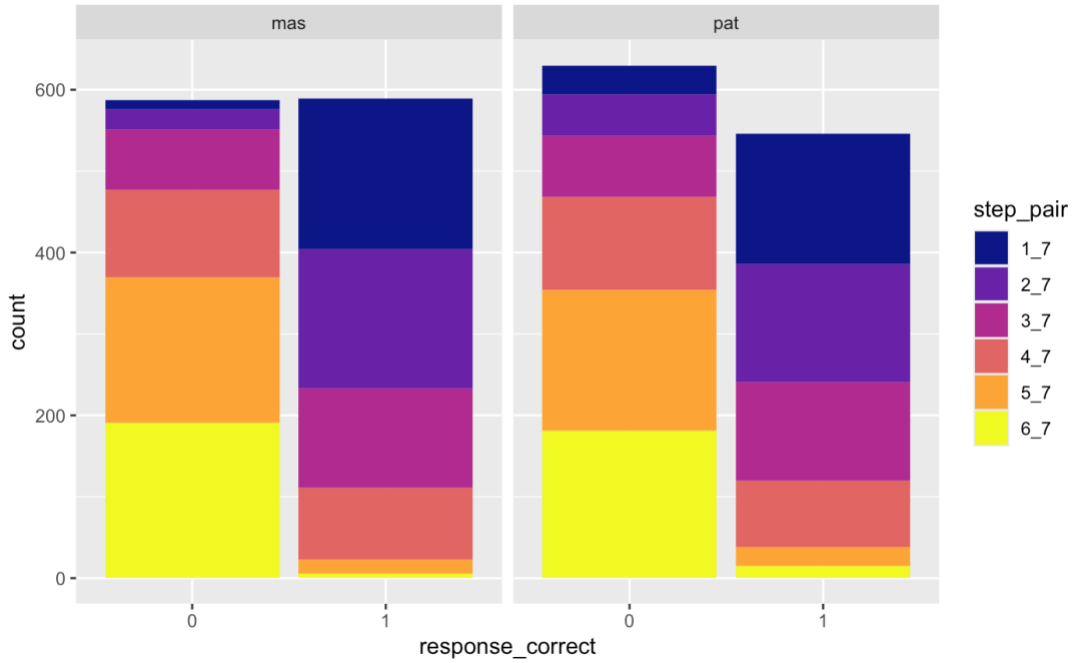


Figure 4.15: Pooled responses by WORD for varying intervals for the [a]/[ɒ]$_A$ vowel continuum. The right-hand column represents correct responses, while the left-hand column represents incorrect responses (i.e., failure to discriminate the interval).

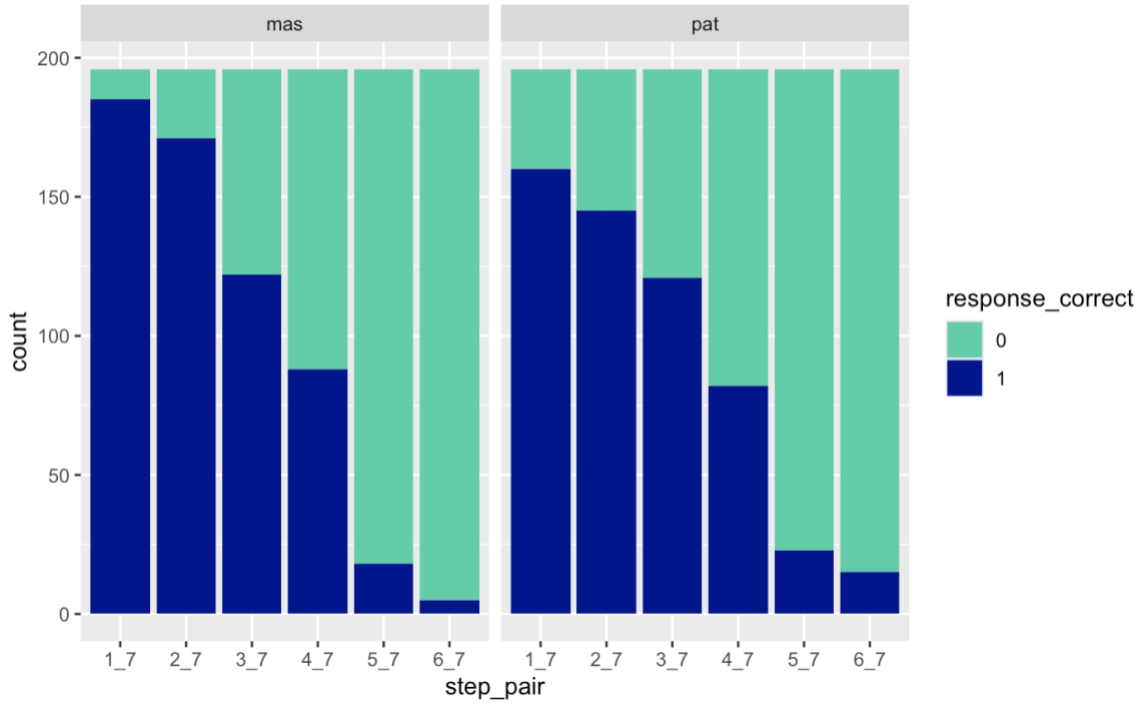


Figure 4.16: Pooled responses by WORD for varying intervals for the [a]/[ɒ]$_A$ vowel continuum. Dark blue represents correct responses, while green represents incorrect responses (i.e., failure to discriminate the interval).

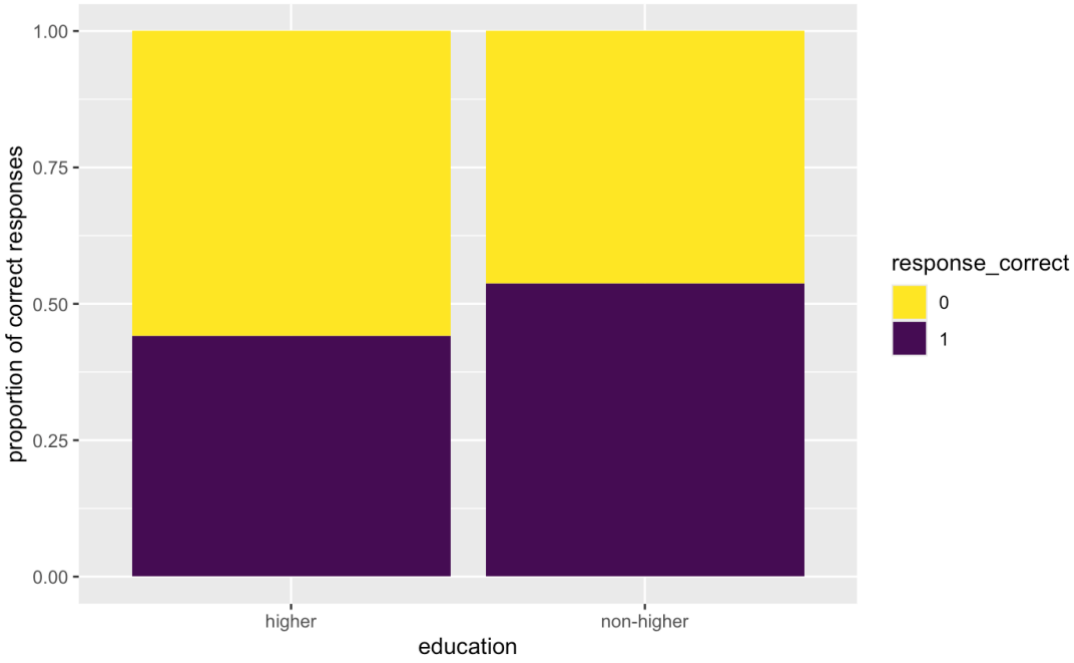


Figure 4.17: Proportion of correct responses by EDUCATION LEVEL for varying intervals for the [a]/[ɒ]$_A$ vowel continuum. Purple represents correct responses, while yellow represents incorrect responses (i.e., failure to discriminate the interval).

4.6.2.2.2 [ɔ]/[ɒ]o

In the mixed-effects logistic regression models predicting responses to varying intervals for the [ɔ]/[ɒ]o continuum, several production and demographic predictors emerged as being potentially relevant for predicting the outcome variable. The simplest model (which included fixed predictors of STEP PAIR, WORD, and their interaction, along with random intercepts for trial order, random intercepts for participants, and random slopes for participants relative to STEP PAIR, WORD, and their interaction) had an AIC of 2345.2. However, a model that also included PREVIOUS (with previous participants having a higher proportion of correct responses) had a slightly better AIC of 2344.7, with this predictor having a sub-significant p-value of 0.108. A model that included both ED-median (calculated from non-Bark-converted data) and Pillai scores (calculated from Bark-converted data, though as noted in Chapter 2, Bark-conversion resulted in only very small changes to Pillai scores likely caused by rounding) also had a similar AIC of 2345.3, and Pillai scores were just barely significant in this model (p=0.050). However, the interpretation of this model was strange, in that its coefficients suggested that Pillai scores and ED-median were related to responses in opposite ways, with larger Pillai scores (less overlap between vowels) being associated with less correct responses, and larger distances (ED-median) between vowels in production being associated with more correct responses. Notably, Pillai scores were not significant (p=0.243; AIC=2345.9) in a model that did not also include ED-median. These circumstances might suggest that the given combination of predictors is leading to overfitting or overinterpretation of noise in the data, as it is not possible to interpret the opposite effects of the two production metrics in a meaningful way (i.e., we cannot say that either more dialect-like or less dialect-like vowels in production correspond with better discrimination, because the coefficients of the two production metrics point to opposite interpretations). Furthermore, examination of the Pillai score data showed that the participant with the lowest accuracy on the AX varying intervals for [ɔ]/[ɒ]o (0% correct responses) also had the very highest Pillai score for this vowel pair. The significant effect found for Pillai scores seems to have been entirely driven by this participant, as omitting them from the model drastically changed the significance of the effect (p=0.386). In light of these issues, the model that included only PREVIOUS (in addition to the core model structure) is presented in **Table 4.40** below.

As **Table 4.40** shows, STEP PAIR had a positive slope, indicating that there were more correct responses as intervals became larger for the reference level *pos.* The significance and

negative coefficient for WORD indicate that accuracy was overall lower for *tsots* than for *pos*, which can also be seen in **Figures 4.18-4.19**. The interaction term is not significant, indicating that the slope of STEP PAIR was not significantly different for the two words.

Table 4.40: Results of mixed-effects logistic regression model predicting responses to varying intervals for [ɔ]/[ɒ]$_O$. The model structure was as follows: **RESPONSE CORRECT ~ STEP PAIR * WORD + PREVIOUS + (STEP PAIR * WORD | participant) + (1 | trial_order)**. The reference level for WORD was *pos*. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places.

Predictor	**estimate**	**p-value**
STEP PAIR	1.535	*<0.001*
WORD (*tsots* vs. *pos)*	-1.96	*<0.001*
PREVIOUS	0.649	0.108
STEP PAIR – WORD interaction	0.228	0.558

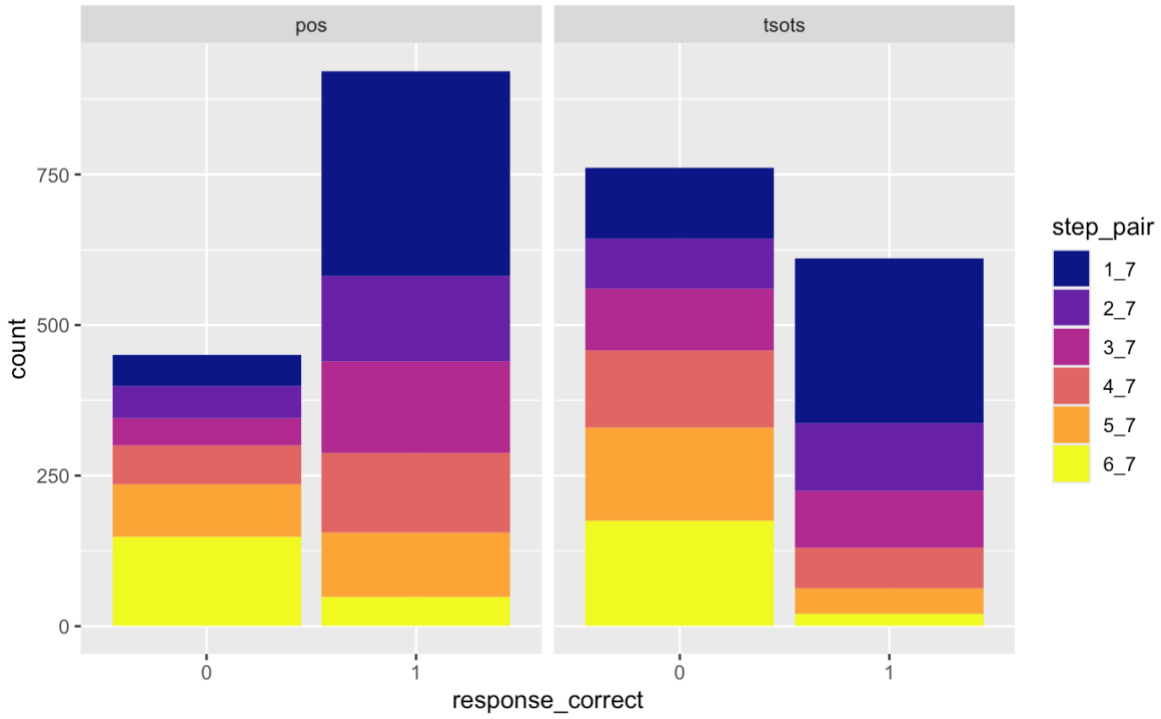


Figure 4.18: Pooled responses by WORD for varying intervals for the [ɔ]/[ɒ]$_O$ vowel continuum. The right-hand column represents correct responses, while the left-hand column represents incorrect responses (i.e., failure to discriminate the interval).

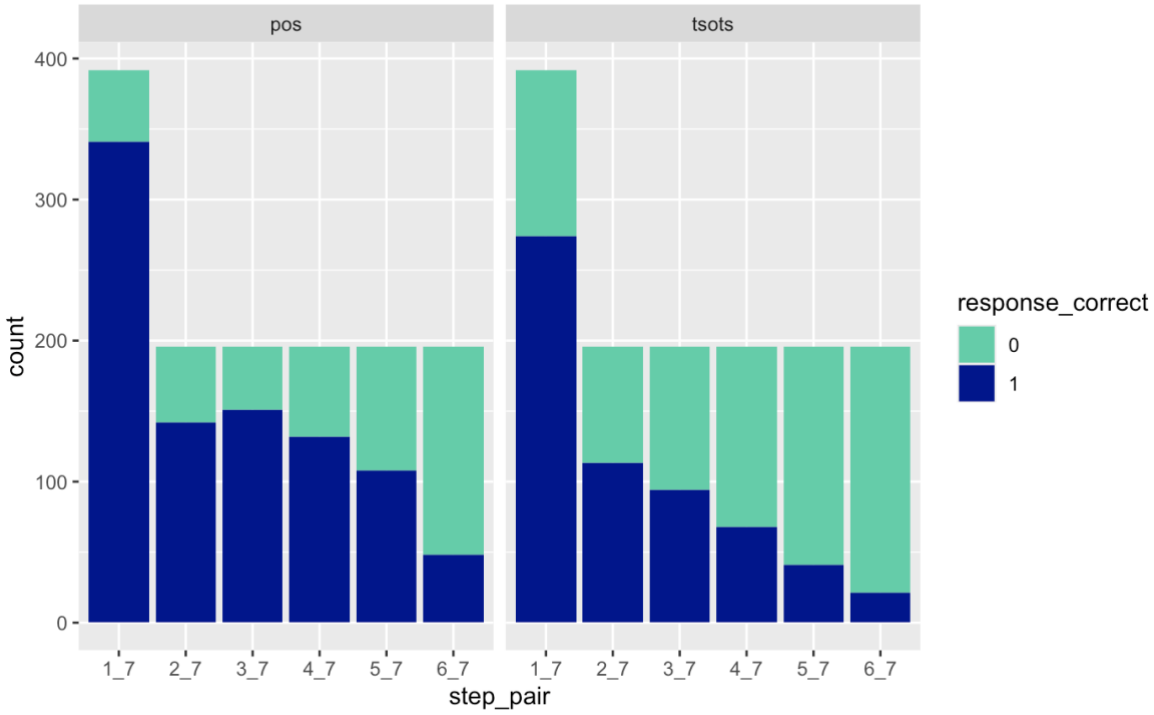


Figure 4.19: Pooled responses by WORD for varying intervals for the [ɔ]/[ɒ]$_O$ vowel continuum. Dark blue represents correct responses, while green represents incorrect responses (i.e., failure to discriminate the interval). Recall that the interval 1-7 had extra repetitions for this vowel pair, as explained in Section 4.3.3.3.

4.6.2.3 Summary of results

As demonstrated by the results presented in Section 4.6.2.1, the logistic regression models predicting responses to consistent intervals found significant differences between responses for the two words representing each vowel pair. Specifically, correct responses were significantly higher for *mas* (compared with *pat*) and for *pos* (compared with *tsots*). Though the differences between the slopes of STEP PAIR for the two words were significant for *mas/pat* and nearly significant for *pos/tsots*, the slopes within each vowel pair had the same directionality (negative slopes for *pat/mas* and positive slopes for *pos/tsots*). Visualizations of raw data showed that the overall response pattern was similar for the two words representing [a]/[ɒ]$_A$ (*mas* and *pat*), in that correct responses dropped off sharply for the interval 5-7, which represented small differences from the SEA end of the continuum. No demographic or production metrics were predictive of correct responses for [a]/[ɒ]$_A$, but EDUCATION LEVEL and BIRTH YEAR were sub-

significant in models predicting [ɔ]/[ɒ]$_O$, with people with higher education and younger people having lower accuracy.

In the logistic regression models predicting correct responses to the intervals of varying size, the slope of STEP PAIR was positive for both words for both vowel pairs, meaning that correct responses generally increased as intervals became larger. For [a]/[ɒ]$_A$, there was no significant difference between the overall responses to the two words (*mas* and *pat*), though their slopes were significantly different, while for [ɔ]/[ɒ]$_O$, there was a significant difference between the overall responses for the two words (*pos* and *tsots*), but no significant difference between their slopes. Production metrics were not predictive of responses for either vowel pair, and the only demographic trends found were a significant effect of EDUCATION LEVEL for [a]/[ɒ]$_A$ (with people with higher education having lower accuracy) and a sub-significant effect of PREVIOUS for [ɔ]/[ɒ]$_O$, in which repeat participants had more correct responses.

4.7 General discussion

4.7.1 Comparison of results for the two vowel pairs

The results of the linear and logistic regression models and data visualizations paint a picture in which responses for the two different vowel pairs in the AX task conform to some expectations but defy others. As noted in Section 4.3.2.1, the acoustic differences in the stimuli were much larger for [a]/[ɒ]$_A$ than for [ɔ]/[ɒ]$_O$. Even considering that the consistent intervals presented in the task were 2-step intervals for [a]/[ɒ]$_A$ and 3-step intervals for [ɔ]/[ɒ]$_O$, the differences presented in the task were always larger for [a]/[ɒ]$_A$. In spite of this, performance was similar across the two vowel pairs, as measured by ACCURACY, CONSISTENT INTERVALS ACCURACY, and *d'*. The linear models involving these outcome variables supported the patterns observed in the data visualizations, finding that there were sometimes significant differences between the members of each vowel pair (e.g., between *pos* and *tsots*) according to which word had larger vs. smaller acoustic differences in the stimuli, but never significant differences between the vowel pairs. It should be recalled that [ɔ]/[ɒ]$_O$ had additional repetitions of the large interval 1-7, which could have inflated metrics of performance on this vowel pair. However, since the pattern of results was similar for ACCURACY, CONSISTENT INTERVALS ACCURACY (which did not include the interval 1-7), and *d'*, these extra repetitions of large intervals likely do not

play a significant role in explaining the unexpectedly good discrimination of [ɔ]/[ɒ]$_O$ stimuli in comparison with the more acoustically distinct [a]/[ɒ]$_A$ stimuli.

In contrast with the other metrics of performance on the AX task, a discrepancy between the two vowel pairs was observed for SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED by each participant. While some participants consistently discriminated the 2-step interval 5-7 for both words representing [ɔ]/[ɒ]$_O$, no participants consistently discriminated this interval for either word representing [a]/[ɒ]$_A$, which is unexpected given the actual acoustic size of the intervals for the two vowel pairs. The visualization of the pooled consistent intervals data (Section 4.6.2.1), together with the results of the logistic regression models for the consistent intervals data, also show differences between the two vowel pairs. These analyses revealed that participants had much more difficulty discriminating the interval 5-7 for [a]/[ɒ]$_A$ than they did discriminating other 2-step intervals for this vowel pair, and that discrimination of 2-step intervals for this vowel pair generally became worse as the intervals moved closer to step 7, with a sharp drop off in accuracy for the interval 5-7. The pattern of responses for the consistent intervals for [ɔ]/[ɒ]$_O$ was not as similar across the two words, with *pos* having a strongly positive slope (better discrimination closer to step 7) and *tsots* having a very weakly positive slope. Some participants reported that stimuli near the right end of the continuum (near step 7) for *pos* sounded like they contained the vowel [u], so the better discrimination toward this end of the continuum for this word could be explained by the continuum having unintentionally included a phonemic category boundary for some listeners. On the other hand, the relatively flat slope for the consistent intervals for *tsots* could indicate that the continuum for this word did not contain any phonologically meaningful landmarks.

Given that the interval 5-7 for the [a]/[ɒ]$_A$ continuum represented small differences from the canonical SEA variant [a], the explanation for the sharp drop-off in accurate discrimination for this interval could be related to the phonological characteristics of this vowel pair. In particular, among all the 2-step differences along the continuum, small differences from the SEA endpoint [a] are most similar in their magnitude and location in the vowel space to the small, potentially coarticulatory differences identified in Yerevan by Seyfarth and Garellek (2018). Thus, in hearing the small differences represented by the interval 5-7 (and also by the interval 6-7, which no participant discriminated consistently), it is possible that listeners attributed the slightly lower F1 and F2 values of the vowel (as compared with the endpoint step 7) to the

consonantal environments of the words *mas* and *pat*. Specifically, as discussed in Chapter 3, nasal contexts can contribute to lowering of both formants (see Carignan 2018), while the plain voiceless initial consonant of the word *pat* lacks the breathiness associated with a fronter [a] in Yerevan Armenian (see Seyfarth & Garellek 2018). Thus, the potential for listeners to perceive steps 6 and 5 as variants that are expected in the given consonantal contexts may have inhibited discrimination of these steps from step 7. In contrast, starting from step 4, the lowering of F1 and F2 (in comparison with step 7) may have moved beyond the range of coarticulatory variants that listeners expect in the given contexts, facilitating discrimination.

If we assume that the [a] end of the continuum aligns with listeners' prototypical representation of the phoneme /a/, then the poor discrimination near this end of the continuum could also perhaps be explained by the "perceptual magnet effect" identified by Kuhl (1991), in which vowels that are closer to listeners' prototypical vowel are more difficult to discriminate. While there are thus multiple potential explanations for the poor discrimination of the pairs 5-7 and 6-7 for this vowel pair, the overall unexpectedly poor discrimination in comparison with discrimination of the less acoustically distinct [ɔ]/[ɒ]$_{O}$ stimuli can likely be explained by the allophonic nature of the [a]/[ɒ]$_{A}$ pair, as previous studies have found perception of allophonic differences to be more difficult than perception of contrastive differences (see Section 3.4.2).

While the tendencies related to performance at different points along the continua discussed in the previous paragraphs pertain to group-level patterns, participants exhibited considerable variation in their individual performance on the consistent intervals. There were differences not only in the locations along the continua where they exhibited better discrimination, but also in the overall shape of their responses. In other words, while for some participants, a clear peak in accuracy could be identified, others had more diffuse areas of higher accuracy, areas of higher accuracy at multiple non-adjacent areas of the continuum, or areas of higher accuracy that were still relatively low (i.e. 25% accuracy, compared to 0% accuracy elsewhere on the continuum). Identifying whether/where each participant had a peak in perceptual accuracy thus proved exceedingly difficult and would have involved a great deal of arbitrary subjective judgements. As such, analysis of the individual-level consistent intervals data was used primarily for within-participant comparisons, comparing individual performance on the AX and dialect categorization tasks. In particular, for each participant it was considered whether their performance on the consistent intervals portion of the AX task (i.e., the location of

their areas of higher accuracy) could explain certain aspects of their performance on the categorization task, as discussed in Chapter 6.

4.7.2 Production metrics and demographic factors

As the analyses above demonstrated, no relationships were found between production metrics (Pillai scores and ED-median) and performance on the AX task. However, there were a few demographic trends that were found across multiple analyses. First, BIRTH YEAR was found to be a significant predictor of multiple performance metrics for both vowel pairs (and a sub-significant predictor of other metrics). Second, EDUCATION LEVEL was found to be a significant predictor for multiple performance metrics for [a]/[ɒ]$_A$, as well as in the varying intervals logistic regression model for this vowel pair, and was also a sub-significant predictor of some other metrics for both vowel pairs. EDUCATION LEVEL and BIRTH YEAR were also sub-significant in the models predicting responses to varying intervals for [ɔ]/[ɒ]$_O$. In all of these cases, younger people showed worse discrimination than older people, and people without higher education had better discrimination than those with higher education (though it should be noted that a sub-significant opposite education trend was found in the consistent intervals logistic regression model for [ɔ]/[ɒ]$_O$).[26] The age effect is difficult to explain, as some previous studies (e.g., Harris et al. 2008) have found that older adults have worse frequency discrimination than younger adults. Though the older participants in the present study (with maximum age of 50) were younger than those who are considered "older" by Harris et al. (2008), some studies (e.g., König 1957) have found worsening pitch discrimination even in adults under age 55. Nevertheless, these studies examined frequency discrimination in pure tones (pitch), rather than discrimination of vowel formant frequencies. It is possible that older participants in the present study were able to discriminate vowels more accurately thanks to their greater accumulated lifetime experience with perceiving language (or with perceiving dialect vowels in particular), especially given that the stimuli were based on natural speech from their own speech community. The finding that participants without higher education showed better vowel discrimination than those with higher education is consistent with some previous findings related to gradience of speech perception. Namely, Kutlu et al. (2024) found that children with more linguistically diverse social environments had more gradient and flexible speech perception, while Kolinsky et al. (2021)

[26] Age and education level were not strongly correlated with one another, with a correlation of less than 0.1.

found that lower literacy was associated with less precise phonemic categories. While it is difficult to say if the different education groups of the present study are comparable with participants in these studies, it is possible that less educated participants have more diverse linguistic environments (e.g., due to spending less time interacting with institutions where SEA is used) and/or have lower literacy (though all participants had completed secondary school and were literate in the general sense of the word). In other words, while sufficient information is not known about the participants to say exactly what (other than their education level itself) could be driving this trend, there are various possible explanations that are in line with previous research. In any case, these demographic trends were relatively weak (compared with the significant differences between the words, step pairs, etc.), and the fact that different metrics (e.g., ACCURACY vs. *d'*) sometimes pointed to different demographic trends means that they should be interpreted with caution.

4.7.3 Methodological considerations

Though the analyses of different individual metrics of performance on the task generally pointed toward similar results in relation to the key predictors of WORD and VOWEL PAIR, the results were not identical across different metrics. For example, the results of models predicting ACCURACY and *d'* differed in terms of the significance of differences between certain pairs of words. Thus, although the results across different metrics in relation to WORD and VOWEL PAIR were generally more consistent and/or had larger effect sizes than the results in relation to demographic factors, the results nonetheless demonstrate that the choice of which metric to use to quantify participants' performance can have significant downstream effects in the analysis. The results presented in this chapter thus suggest that different methods should be explored and compared, especially in cases where there is no a priori reason for choosing one method over another, or where there exist arguments in favor of both methods. Though previous authors have also called for methodological transparency in linguistics (e.g., Bochynska et al. 2023), the results presented in this chapter underscore the particular importance of transparency about *why* certain methods are chosen (and the implications of these choices).

A similar conclusion can be drawn from comparing the models that separated the two vowel pairs with those that pooled them. For example, demographic effects that were significant in separate models often did not correspond to significant interactions in the pooled models, even if a tendency toward significance was noted in the pooled models. This suggests that

methodological decisions can have a substantial effect on which underlying trends in a data set are noticed and deemed to be significant, underscoring the need for caution in relying on arbitrary statistical significance thresholds, particularly in analyses that are largely exploratory and that involve weak trends. In other words, the mixed results of the analyses presented in this chapter, in which certain trends repeatedly emerged across different analyses but varied in their statistical significance, indicate a need for transparency about how methodological decisions affect statistical results, as well as a general need for decreased reliance on "binary decision procedures" (see Etemady & Roettger 2025: 12).

Chapter 5 Dialect Categorization Task: Social Salience of Vowels may Facilitate Categorization

5.1 Chapter summary

This chapter presents the goals, expectations, methods, and results for the dialect categorization task. The chapter first summarizes how the task investigates participants' perception of the dialectal association of the [a]/[ɒ]$_A$ and [ɔ]/[ɒ]$_O$ vowel continua, highlighting how expectations about responses to the two pairs differ in connection with their differing social salience. After presenting the task design and information about the stimuli, the chapter then discusses statistical methods used for group-level analysis (mixed-effects logistic regression models), as well as the calculation of individual-level metrics of performance and analysis of these metrics using mixed-effects linear and logistic regression models. Discussion of the results focuses on differences between responses for the two vowel pairs. The most important finding was that the [a]/[ɒ]$_A$ stimuli were categorized in a more expected manner than [ɔ]/[ɒ]$_O$ (i.e., with more "Gavaṛ" responses toward the dialect-like end of the continuum and more "not Gavaṛ" responses toward the SEA-like end of the continuum) according to multiple different analyses. This was an expected outcome based on the social salience of [a]/[ɒ]$_A$ in comparison with [ɔ]/[ɒ]$_O$. Analyses of multiple individual-level metrics also suggested that younger people and men responded in a less expected manner than older people and women.

5.2 Goals and expectations of the dialect categorization task

This chapter asks which steps along continua from [a] to [ɒ]$_A$ and [ɔ] to [ɒ]$_O$ will be associated with Gavaṛ's dialect by local listeners, and which steps will not be associated with Gavaṛ's dialect. The core goal of the task is to investigate how response patterns vary across these two vowel pairs and how the social and phonological characteristics of the vowels could explain such patterns. Importantly, despite the differing phonological characteristics of the two vowel pairs ([a]/[ɒ]$_A$ being allophonic and [ɔ]/[ɒ]$_O$ being contrastive), the dialect categorization

task examines a socially mediated, non-phonemic type of categorization in the case of both vowel pairs, as the stimuli do not consist of minimal pairs for either vowel pair.

The two endpoints of both the [a]/[ɒ]$_A$ and [ɔ]/[ɒ]$_O$ continua represent dialect (step 1) and SEA (step 7) pronunciations, with [ɒ]$_A$ and [ɒ]$_O$ representing dialect pronunciations and [a] and [ɔ] representing SEA pronunciations. When presented with these most extreme points from each continuum, it was predicted that listeners would thus classify [ɒ]$_A$ and [ɒ]$_O$ as characteristic of Gavaṛ's dialect, and [a] and [ɔ] as not characteristic of Gavaṛ's dialect. Although the sounds [a] and [ɔ] are also found in the sound system of the dialect, the perception stimuli consisted of words in which [ɒ]$_A$ and [ɒ]$_O$ are expected in the dialect, while [a] and [ɔ] are expected in SEA.

Because the [a]/[ɒ]$_A$ allophony is more socially salient in its association with the dialect according to the prior literature (e.g., Katvalyan 2016), it was expected that listeners might categorize the endpoints of this continuum more clearly than in the case of the [ɔ]/[ɒ]$_O$ continuum. There is evidence that the [ɒ]$_A$ vowel is not only more salient, but is also a feature that distinguishes Gavaṛ's dialect even from its current close neighbors. Katvalyan (2025: 38) writes that this vowel is used in some varieties of the Mush and Maku dialects spoken elsewhere in Geghark^{h}unik^h province, but its use in these varieties is "partial," as opposed to its "regular" (i.e., phonologically regular) and "stable" use in the Bayazet dialect. In contrast, Katvalyan (2016: 91-92) writes that many of the former neighboring dialects of Gavaṛ (including some varieties of the Ararat dialect group, which still neighbors it today)[27] have phenomena similar to the [ɔ]/[ɒ]$_O$ contrast, and Katvalyan (2025: 66) notes that such phenomena are also currently widespread in the dialects of Geghark^{h}unik^h province. Katvalyan (2016: 68) nevertheless suggests that the phonetic manifestation of the sound equivalent to [ɒ]$_O$ is not the same in other dialects, so it is possible that the manifestation of this phoneme specifically as a sound that is phonetically similar to [ɒ]$_A$ could be perceived as locally specific to Gavaṛ. Thus, in general, despite the phonetic similarity of [ɒ]$_A$ and [ɒ]$_O$,[28] differences in their salience and local specificity could lead to differences in how they are categorized.

[27] The population of Gavaṛ is said to be largely descended from 19th-century migrants from the city of Bayazet (located in present-day Turkey) (Katvalyan 2016: 31). In connection with this, Katvalyan (2016) refers to the city's dialect as the "Bayazet dialect," and gives detailed comparisons of its features with the features of those dialects that neighbored it in its original location (many of which no longer neighbor it today and/or are extinct today).

[28] Although I transcribe these two sounds with the same IPA symbol based on Katvalyan's assertion that they are pronounced identically, my participants generally did not have fully overlapping distributions for these two sounds even in cases where they maintained large distinctions from [a] and [ɔ]. There are also other claims (e.g., Khachaturian 1992) that the two vowels are not pronounced completely identically.

If the endpoint stimuli were to be clearly categorized in the expected manner, it was hypothesized that participants would transition from associating the pronunciations with Gavaṛ's dialect to not associating them with Gavaṛ's dialect at some intermediate point along each continuum. Though there were no specific predictions about where on each continuum listeners (as a group or individually) would cross over, the crossover might be expected to be more clear for the socially salient [a]/[ɒ]$_A$ pair. In general, the salience of this vowel pair could lead listeners to categorize it in a more expected way (i.e., with clearer endpoints corresponding to "Gavaṛ" responses for the dialect-like endpoint and "not-Gavaṛ" responses for the SEA-like endpoint, and with a clearer crossover [i.e., more categorical response pattern]). In general, because the two vowel pairs are separate phenomena occupying different portions of the vowel space, it was expected that crossover points would likely differ across the two vowel pairs (in terms of their locations along the continuum). However, it was also expected that such differences might relate to performance on the AX task in some way (e.g., with crossover points corresponding to peaks in accuracy on the AX task), or that other aspects of performance on the two tasks might be related, as explored in the analyses presented in this chapter and in more detail in Chapter 6.

Although the main goal of this chapter is to compare responses to the dialect categorization task across the two vowel pairs, the analyses presented in this chapter also include some exploration of the potential role of listeners' demographic characteristics in perception, of relationships between perception and production, as well as of relationships between performance on the two perception tasks. Results in relation to the latter two topics will thus be mentioned in this chapter, but they will also be further discussed in Chapter 6, where additional qualitative (i.e., not involving inferential statistics) analyses of these topics will be presented.

5.3 Dialect categorization task – stimuli, task structure, and data collection

5.3.1 Types of stimuli

The test stimuli for this task consisted of the same type of stimuli used in the AX task – namely, words representing various steps along the [a]/[ɒ]$_A$ and [ɔ]/[ɒ]$_O$ continua. Participants heard one word at a time with vowels from these continua – for example, [pɒt] – and chose if the pronunciation is characteristic of Gavaṛ's dialect or not. The use of the same stimuli in the AX and dialect categorization tasks was intended to facilitate the comparison of the results of the two tasks. For example, a participant might judge two steps from the [a]/[ɒ]$_A$ continuum of a certain

lexical item as different in the AX task but categorize them together in the geographic categorization task, which would tell us that the participant can discriminate within-category phonetic differences that do not necessarily correspond to the phonetic boundaries that they ascribe to dialect and non-dialect variants. On the other hand, a participant could fail to discriminate small acoustic differences in the AX task, but show clear categorization of those same stimuli in the second task, which would show that the social prompting of the second task perhaps allowed them to attend to more subtle phonetic differences. It could also be found that participants' dialect vs. non-dialect perceptual category boundaries align with some aspects of performance on the AX task, such as areas of higher accuracy in the consistent intervals data, or the SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED from the varying intervals data. In addition to the stimuli that were the exact same across the two tasks, one additional word each from the [a]/[ɒ]$_{A}$ and [ɔ]/[ɒ]$_{O}$ continua was included in the dialect categorization task, so that it could be determined if participants performed dramatically differently on stimuli that they had vs. had not already encountered.

The training portion used three types of stimuli to familiarize participants with the task. First, in order to once again make participants aware that the task involves categorization of words that differ in the vowel, the training portion presented some [a]/[æ] stimuli. Participants heard, for example, the pronunciation [d͡zækʰ] and had to choose if this pronunciation is characteristic of Gavaṛ's dialect or not. Throughout the training, they heard various steps from some of the [a]/[æ] continua. In general, the same [a]/[æ] stimuli were used in this task as in the AX task.

As discussed in Chapter 4, though some participants of the 2022 study did associate [æ] with the local dialect in their commentary, this vowel is widespread in dialects of Armenian, including those of Geghark^{h}unik^{h} province. (Though the [ɔ]/[ɒ]$_{O}$ phenomenon used as test stimuli may also not be locally specific in terms of phonology, there is at least reason to believe that its phonetic manifestation is locally specific, as discussed above in Section 5.2). Thus, because it was unclear whether the [æ] vowel would produce a strong enough association with Gavaṛ, the training for this task also included some other phenomena that might have been more likely to be perceived as locally specific. One of these phenomena was words in which SEA [h] is pronounced with [χ] in the dialect – for example [hat͡sʰ]/[χat͡sʰ] ('bread') and [hars]/[χars] ('bride'). Pairs of stimuli were created using the interview data of the same talker who provided

the other stimuli, by splicing initial [h] into words originally pronounced with [χ]. Details about the creation of these stimuli are found below in Section 5.3.2.2. The use of initial [χ] in words where SEA uses [h] is a highly salient and still widely used local dialect feature based on the 2022 interview data. Although Katvalyan (2025: 97) writes that this feature is found in other varieties spoken in Geghark^{h}unik^{h} province, he notes that it is "fully manifested" in the Gavaṛ region, meaning it may be more strongly associated with Gavaṛ's dialect than with other dialects. Thus, presenting listeners with words like [χa͡ts^{h}] was likely to prompt categorization of the tokens with Gavaṛ's dialect. The order of these stimuli throughout the task was controlled, such that listeners heard and categorized, for example, the dialect variant [χa͡ts^{h}] prior to hearing and categorizing the supralocal variant [ha͡ts^{h}]. Because some residents of Gavaṛ use the supralocal pronunciation [ha͡ts^{h}], presenting the stimulus [χa͡ts^{h}] prior to the stimulus [ha͡ts^{h}] was intended to help participants to better understand that they should categorize obviously local pronunciations as being characteristic of Gavaṛ's dialect and supralocal pronunciations as being not characteristic of Gavaṛ's dialect.

In case the phonetic differences presented by the [a]/[æ] and [h]/[χ] stimuli were not sufficient to familiarize participants with the nature of the task, the training also included lexical items that were either neutral or strongly associated with Gavaṛ's dialect. Although it would have been preferable to present corresponding pairs of such stimuli, such that participants would hear, for example, both the dialect variant [hæʃk] ('eye') and the SEA variant [a͡tʃk^{h}] ('eye'), no such pairs could be obtained from the talker's interview data. Thus, this subset of the stimuli simply consisted of some words that are strongly associated with Gavaṛ's dialect, and other words (with different meanings) that are found in SEA and are likely to also be used by residents of Gavaṛ, such that they have a neutral dialect association. It was hypothesized that participants might mostly choose "not Gavaṛ" when presented with these neutral words, or that they would perform at or near chance, since the dialect association of these words is ambiguous. In contrast, it was expected that participants would choose "Gavaṛ" when presented with the dialect words. In the training portion of the task, some of the dialect words were presented immediately before some of the neutral words, such that participants would understand that they should judge the neutral words against the backdrop of the more obviously dialectal words. Details about the procedure used to create these stimuli are found below in Section 5.3.2.2.

The [a]/[æ] stimuli, [h]/[χ] stimuli, and lexical stimuli were also all used during the test portion of the task – the (potentially less locally specific) [a]/[æ] stimuli as filler and the other two types as control conditions used to check if participants understood the task. Words with initial [χ] (as opposed to SEA [h]) were expected to be categorized consistently as Gavaṛ dialect pronunciations due to their high salience and due to the stronger local specificity and categoricity of this feature (in contrast to the more gradient nature of vowels). Participants' consistent categorization of these stimuli in the expected manner would demonstrate that they had correctly learned how to perform the task, lending credibility to the other results. In case the [h]/[χ] alternation turned out to not be salient enough to trigger the expected categorization, the dialect and neutral lexical items were intended to provide an additional means of verifying that participants had understood the task.

5.3.2 Creation of stimuli

5.3.2.1 New test stimuli

As noted above, two additional words (one representing each vowel continuum) were included in this task. These words were [marʈʰ]/[mɒɾʈʰ] ('person') for the [a]/[ɒ]$_{A}$ continuum and [χɔz]/[χɒz] ('pig') for the [ɔ]/[ɒ]$_{O}$ continuum. The remaining test stimuli for these two vowel pairs were identical to those used in the AX task (see Chapter 4), as were the training/filler stimuli representing the [a]/[æ] continuum.

The stimuli were created using the same procedure described above in Section 4.3.2.1, including the modification of vowel duration, preceding and following silence, and intensity. **Table 5.1** below reports the maximum formant value setting used in creating each continuum. **Tables 5.2** and **5.3** report the mean F1 and F2 values in Hz and Bark for each step of each continuum. **Table 5.4** reports the F1 and F2 differences in Bark between consecutive steps of each continuum. All Bark values were obtained using the *barktools* package (Stanley 2020).

Notably, the difference in F2 between the endpoints of the [χɔz]/[χɒz] ('pig') continuum is in the opposite direction of what is expected for this vowel pair, thus differing from the continua of the two words used to represent this vowel pair in the AX task. Several factors contributed to the decision to nevertheless use this word in the dialect categorization task. Firstly, in the two continua representing this vowel pair in the AX task, the F1 difference between the endpoints is much larger than the F2 difference, suggesting that the difference between the

dialect and neutral variants is largely driven by F1, with F2 differences possibly being less important for the production/perception of this vowel pair. Secondly, although the F2 difference for [χɔz]/[χɒz] ('pig') is not in the expected direction, it is very small in magnitude, being substantially smaller than the already small F2 differences found in the two words used in the AX task. Although the F1 difference for [χɔz]/[χɒz] ('pig') is also slightly smaller than the F1 differences of the two words used in the AX task, the three continua sounded subjectively similar to one another in terms of the nature and magnitude of the step differences. Though only one pilot participant responded in the expected manner (i.e., clear crossover point and expected classification of the endpoints) for this vowel pair, this pilot participant responded in a similar manner to all three words representing this vowel pair, suggesting that these small differences in step size/direction are not perceptually significant. Crucially, there were also no other more suitable continua that could be derived from the talker's previously recorded speech.

Table 5.1: Maximum formant value setting used in the creation of each continuum with Winn's Praat script.

word	**maximum formant value setting**
[maɾtʰ]/[mɒɾtʰ]	2000 Hz
[χɔt]/[χɒt]	2000 Hz

Table 5.2: Mean F1 values in Hz and Bark for each step of the [maɾtʰ]/[mɒɾtʰ] and [χɔt]/[χɒt] continua.

word	**step 1**		**step 2**		**step 3**		**step 4**		**step 5**		**step 6**		**step 7**	
	Hz	**Bark**	**Hz**	**Bark**	**Hz**	**Bark**	**Hz**	**Bark**	**Hz**	**Bark**	**Hz**	**Bark**	**Hz**	**Bark**
[maɾtʰ]/ [mɒɾtʰ]	487	4.81	522	5.11	558	5.41	595	5.71	633	6.01	673	6.32	714	6.63
[χɔz]/ [χɒz]	454	4.51	438	4.37	423	4.23	407	4.08	392	3.94	377	3.80	363	3.66

Table 5.3: Mean F2 values in Hz and Bark for each step of the [maɾtʰ]/[mɒɾtʰ] and [χɔt]/[χɒt] continua.

word	**step 1**		**step 2**		**step 3**		**step 4**		**step 5**		**step 6**		**step 7**	
	Hz	**Bark**	**Hz**	**Bark**	**Hz**	**Bark**	**Hz**	**Bark**	**Hz**	**Bark**	**Hz**	**Bark**	**Hz**	**Bark**

[marʇʰ]/ [mɒrʇʰ]	785	7.14	835	7.48	887	7.82	942	8.17	998	8.51	1057	8.86	1118	9.21
[χɔz]/ [χɒz]	876	7.75	879	7.77	883	7.80	887	7.82	890	7.84	894	7.87	898	7.89

Table 5.4: F1 and F2 differences (Bark) between consecutive continuum steps for each word.

word	F1 step (Bark)	F2 step (Bark)
[marʇʰ]/[mɒrʇʰ]	0.30	0.34
[χɔz]/[χɒz]	-0.14	0.02

5.3.2.2 New training/control stimuli

The words making up the [h]/[χ] stimuli are presented below in **Table 5.5**. These stimuli were created from the speech of the same talker who provided all the other stimuli. All four words with initial [χ] came from the talker's actual pronunciations, while the variants with initial [h] were created by splicing in an instance of [h] taken from elsewhere in the talker's speech. For the three words in which [χ] is followed by [a], this [h] came from an instance of the word *ha* ('yes'), while for the word *χing*, it came from the word *himi* ('now').

Prior to the creation of the variants starting with [h] via splicing, the average intensity of the two words within each pair was normalized to match the member of the pair with lower intensity using Praat Vocal Toolkit (Corretge 2012-2024). The boundaries of the [χ] and [h] were marked in each audio file using Praat text grids. The [χ] and preceding silence were removed from the files, before the [h] from the corresponding vowel context was spliced in, along with 100 ms of preceding silence. Following Wagner et al. (2006) and Steinberg et al. (2012), it was ensured that the boundaries of all spliced segments were located at zero-crossings (moments when the amplitude of the sound wave was zero), in order to prevent discontinuities from being introduced into the stimulus.

The splicing procedure was slightly more complex for the word *χat͡sʰ,* which came from connected speech. Prior to the procedure described above, approximately 100 ms of silence (with boundaries at zero-crossings) was copied from nearby in the interview and pasted before and after this word. The procedure for creating the [h] variant of the word *χing* was also slightly more complex, since the [h] obtained from the word *himi* was much shorter than the original [χ] in

χing. To make the two members of the pair more comparable, part of the [h] was copied and pasted multiple times, until its duration was similar to that of the original [χ]. 100 ms of silence from before the word *himi* was then pasted at the beginning of the file. Extra silence was also trimmed from the ends of all stimuli files, such that 100 ms of silence remained.

Table 5.5: Words used as a control condition in the dialect categorization task, in which SEA [h] corresponds with [χ] in Gavaṛ dialect pronunciations.

SEA pronunciation (created by splicing)	**Gavaṛ dialect pronunciation (talker's original pronunciation)**	**Meaning**	**Source of spliced [h]**
[hat͡sʰ]	[χat͡sʰ]	'bread'	[ha] ('yes')
[haɾs]	[χaɾs]	'bride'	[ha] ('yes')
[ha]	[χa]	'yes'	[ha] ('yes')
[hing]	[χing]	'five'	[himi] ('now')

The lexical stimuli, representing words either strongly associated with Gavaṛ's dialect or having a neutral dialect association, were gathered from the previous interview data of the same talker who provided the basis for all the other stimuli. The stimuli were trimmed such that 100 ms of silence preceded and followed each word. For those words that came from connected speech, the procedure described above for the word *χat͡sʰ* was used to insert silence before and/or after the word. The stimuli used for this condition are summarized below in **Table 5.6**.

Table 5.6: Lexical items representing words strongly associated with Gavaṛ's dialect and words with a neutral dialect association.

Dialect words	**Neutral words**
[t͡ʃʰænæ] ('jaw')	[kʰitʰ] ('nose')
[dæɾgæχ] ('gate')	[kʰit͡ʃʰ] ('little')
[dɛʃæk] ('mattress')	[tʰutʰ] ('mulberry')
[hæʃk] ('eye')	[tʰiv] ('number')
[hangat͡ʃ] ('ear')	[tʰuz] ('fig')

5.3.3 Task structure

5.3.3.1 Training portion

The training portion of this task consisted of 24 trials. A summary of the training stimuli is presented below in **Table 5.7**. The order of the stimuli previews the ordering structure used throughout the task. Words with varying vowels were presented first via the continuum endpoints (steps 1 and 7), and then via more intermediate steps, with progressively smaller intervals between adjacent steps (see the stimuli for [daɾd]/[dæɾd] or [d͡zak^{h}]/[d͡zæk^{h}] in **Table 5.7** below). This ordering was intended to orient the listener regarding the range of vowels that they would hear and to make the task easier by allowing the listener to compare intermediate vowel steps (presented later) to the most extreme pronunciations (presented first). In the very first presentation of the vowel stimuli in the training portion, the dialect variant (step 1) preceded the neutral variant (step 7) for the same reasoning described above for the consonant stimuli. Like in the test portion of the task, the consonant and lexical stimuli in the training portion were interspersed between groups of vowel stimuli. The consonant ([h]/[χ]) stimuli were always (in both the training and test portions) presented one after another, and in the order shown in **Table 5.7**. In other words, they were always presented in pairs, with the dialect ([χ]) variant presented before the neutral ([h]) variant. It was determined via piloting that maintaining this ordering would prevent the task from being too difficult, as factors such as background noise could affect the perception of the initial fricatives in these stimuli, but allowing the listener to directly compare them with one another could mitigate this issue. In contrast, the lexical stimuli were not always presented in adjacent pairs with the dialect member preceding the neutral member. However, the first and third pairs of these stimuli presented in the training portion followed this order (see **Table 5.7**).

Table 5.7: Training stimuli for the dialect categorization task. This table reflects the actual number and order of stimuli trials (no trials were repeated during the training portion).

Presentation order	**Stimulus**	**Type**
1	[χing] (‘five’)	Dialect [h]/[χ] word
2	[hing] (‘five’)	SEA [h]/[χ] word
3	[daɾd]/[dærd] – step 1	[a]/[æ] word

4	[daɹd]/[dærd] – step 7	[a]/[æ] word
5	[daɹd]/[dærd] – step 4	[a]/[æ] word
6	[hæʃk] (‘eye’)	Dialect word
7	[k^{h}it͡ʃ^{h}] (‘little’)	SEA/neutral word
8	[d͡zak^{h}]/[d͡zæk^{h}] – step 1	[a]/[æ] word
9	[d͡zak^{h}]/[d͡zæk^{h}] – step 7	[a]/[æ] word
10	[d͡zak^{h}]/[d͡zæk^{h}] – step 4	[a]/[æ] word
11	[t^{h}iv] (‘number’)	SEA/neutral word
12	[t͡ʃ^{h}ænæ] (‘jaw’)	Dialect word
13	[daɹd]/[dærd] – step 2	[a]/[æ] word
14	[daɹd]/[dærd] – step 6	[a]/[æ] word
15	[daɹd]/[dærd] – step 3	[a]/[æ] word
16	[daɹd]/[dærd] – step 5	[a]/[æ] word
17	[χa] (‘yes’)	Dialect [h]/[χ] word
18	[ha] (‘yes’)	SEA [h]/[χ] word
19	[d͡zak^{h}]/[d͡zæk^{h}] – step 2	[a]/[æ] word
20	[d͡zak^{h}]/[d͡zæk^{h}] – step 6	[a]/[æ] word
21	[d͡zak^{h}]/[d͡zæk^{h}] – step 3	[a]/[æ] word
22	[d͡zak^{h}]/[d͡zæk^{h}] – step 5	[a]/[æ] word
23	[hangat͡ʃ] (‘ear’)	Dialect word
24	[k^{h}it] (‘nose’)	SEA/neutral word

5.3.3.2 Test portion

The test portion consisted of 270 trials (6 blocks of 45 trials), meaning that the entire task (including the 24 training trials) consisted of 294 trials. The test portion was blocked by word – i.e., each of the words representing the [a]/[ɒ]$_{A}$ and [ɔ]/[ɒ]$_{O}$ vowel pairs was separated into its own block. Based on preliminary piloting, it was determined that an unblocked task would be too difficult, as categorizing different vowel continua steps is difficult in isolation from any context

or point of comparison. In order to orient the listener, the first two steps presented for the target word in each block were the endpoints, followed by more intermediate steps and by smaller intervals between adjacent stimuli. The first three presentations of stimuli for each test word were step 7 (most extreme neutral pronunciation) – step 1 (most extreme dialect pronunciation) – step 4 (midpoint pronunciation). This order (which is opposite of the order chosen for the presentation of the [h]/[χ] stimuli and the [a]/[æ] stimuli) was chosen because it was hypothesized that hearing the neutral pronunciation first would help the listener to 1) recognize what word they were hearing and 2) recognize the following stimulus as an extreme dialectal pronunciation of the same word. Because the [a]/[ɒ]$_{A}$ and [ɔ]/[ɒ]$_{O}$ pairs (unlike the [a]/[æ] contrast) are fully or partially allophonic, it was hypothesized that the continua endpoints might be more difficult to recognize as categorically different. This could have been especially problematic for categorizing the dialectal pronunciations, which may not have been perceived as extreme in terms of their roundedness, openness, etc., without an immediate frame of reference to draw attention to these qualities and to the relevant lexical item in general. It was thus predicted that participants might struggle to categorize the dialect variants without immediate prior exposure to the more neutral variants. After this initial presentation of the endpoint and midpoint stimuli, the order of the stimuli varied throughout the block – for example, participants initially heard 7-1-4, but later in the block heard 1-7-4.

Table 5.8 below shows a breakdown of the amounts of different types of stimuli in the test portion of the task. **Figure 5.1** shows an abstract template of the block structure, while **Figure 5.2** shows the actual structure of each block. The task took approximately 20 minutes to complete. The test portion of the task included two 2-minute breaks after the second and fourth blocks, while the remaining blocks were followed by a screen that said "press the spacebar to continue." The breaks were automatically timed, but when the timer ran out, a screen was displayed saying "press the spacebar to continue," so that participants could initiate the start of the next section themselves. All sections of the experiment displayed a progress bar that showed participants the portion of the current section remaining until the next break. Each block of the test portion started and ended with stimuli from the lexical control condition, as it was determined that these stimuli demanded the least careful attention/effort in their categorization, and thus would be least affected by factors such as participants being inattentive at the beginning of a new section. The order of the blocks was the same for all participants, such that any

potential influence of the order would be the same for all participants. This strategy was chosen because of the relatively small number of participants, and because of the possibility of including stimuli order as a predictor in statistical analyses (see Section 5.5). Blocks representing the [a]/[ɒ]$_A$ and [ɔ]/[ɒ]$_O$ vowel pairs were presented in alternating order, and the blocks presenting the new lexical items that were not used in the AX task were presented last (see **Figure 5.2** for the block order). This order of the blocks was chosen so that participants would hear and categorize the words that they had already heard in the AX task without any possible influence from the two new test words, making the results of the two tasks as comparable as possible for those words included in both tasks.

The [a]/[ɒ]$_A$ and [ɔ]/[ɒ]$_O$ vowel pairs were each represented by 3 words (2 words for each pair from the AX task, plus one new word for each pair), such that 6 words with 7 continuum steps each were included in the task. Each step of each word was repeated 4 times, equaling a total of 168 test trials.

Each of the four [h]/[χ] control words had two variants, which were each repeated 5 times for a total of 40 trials. As in the training portion, these words were always presented in pairs with the dialect ([χ]) variant being presented first. The 5 dialect and 5 neutral control words were each repeated twice, for a total of 20 trials. These words were sometimes presented adjacently in pairs, and sometimes presented separately (see **Figure 5.2**). The 7 steps of each of the 4 [a]/[æ] filler words were either presented once or twice – two of the words were presented in two blocks each, while the other two were only presented in one block each (the reason for this was to have enough filler stimuli to make the structure of each block uniform). The total number of [a]/[æ] trials was 42. The control and filler trials together thus totaled 102 trials.

Table 5.8: Summary of the amounts of different types of stimuli in the test portion of the dialect categorization task.

	Test ([a]/[ɒ]$_A$ and [ɔ]/[ɒ]$_O$)	**[a]/[æ]**	**[h]/[χ]**	**Lexical stimuli**
Number of words	6	4	4	10
Number of variants per word (continuum steps or variant pronunciations)	7	7	2	n/a

Repetitions of each variant	4	1 (for two words) 2 (for two words)	5	2
Total trials	168	42	40	20

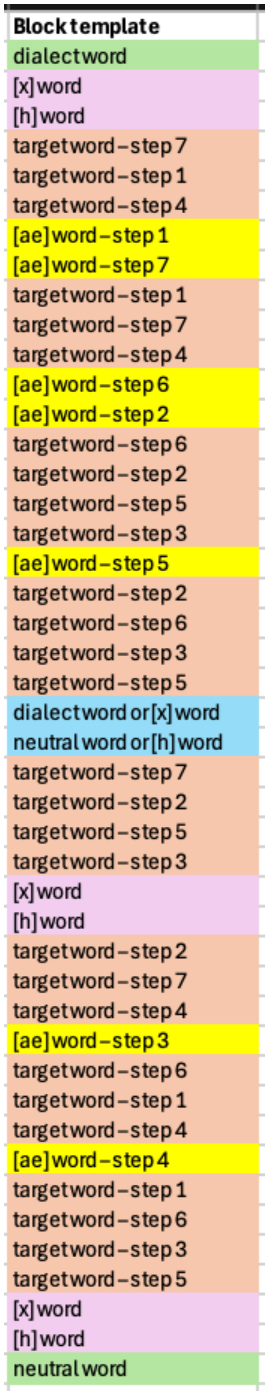


Figure 5.1: Template of the blocks of the dialect categorization task. Pink represents [h]/[χ] stimuli, orange represents test stimuli, green represents lexical stimuli, and yellow represents [a]/[æ] stimuli. Blue represents cases in which some blocks had lexical stimuli and some blocks had [h]/[χ] stimuli.

Block1	Block2	Block3	Block4	Block5	Block6
chaenae.wav	daergaex.wav	deshaek.wav	haeshk.wav	hangach.wav	chaenae.wav
xa.wav	xing.wav	xats.wav	xars.wav	xa.wav	xing.wav
ha.wav	hing.wav	hats.wav	hars.wav	ha.wav	hing.wav
pat_step_7.wav	pos_step_7.wav	mas_step_7.wav	tsots_step_7.wav	mard_step_7.wav	xoz_step_7.wav
pat_step_1.wav	pos_step_1.wav	mas_step_1.wav	tsots_step_1.wav	mard_step_1.wav	xoz_step_1.wav
pat_step_4.wav	pos_step_4.wav	mas_step_4.wav	tsots_step_4.wav	mard_step_4.wav	xoz_step_4.wav
daerd_step_1.wav	dzaek_step_1.wav	naeft_step_1.wav	tat_step_7.wav	daerd_step_1.wav	dzaek_step_1.wav
daerd_step_7.wav	dzaek_step_7.wav	naeft_step_7.wav	tat_step_1.wav	daerd_step_7.wav	dzaek_step_7.wav
pat_step_1.wav	pos_step_1.wav	mas_step_1.wav	tsots_step_1.wav	mard_step_1.wav	xoz_step_1.wav
pat_step_7.wav	pos_step_7.wav	mas_step_7.wav	tsots_step_7.wav	mard_step_7.wav	xoz_step_7.wav
pat_step_4.wav	pos_step_4.wav	mas_step_4.wav	tsots_step_4.wav	mard_step_4.wav	xoz_step_4.wav
daerd_step_6.wav	dzaek_step_6.wav	naeft_step_6.wav	tat_step_2.wav	daerd_step_6.wav	dzaek_step_6.wav
daerd_step_2.wav	dzaek_step_2.wav	naeft_step_2.wav	tat_step_6.wav	daerd_step_2.wav	dzaek_step_2.wav
pat_step_6.wav	pos_step_6.wav	mas_step_6.wav	tsots_step_6.wav	mard_step_6.wav	xoz_step_6.wav
pat_step_2.wav	pos_step_2.wav	mas_step_2.wav	tsots_step_2.wav	mard_step_2.wav	xoz_step_2.wav
pat_step_5.wav	pos_step_5.wav	mas_step_5.wav	tsots_step_5.wav	mard_step_5.wav	xoz_step_5.wav
pat_step_3.wav	pos_step_3.wav	mas_step_3.wav	tsots_step_3.wav	mard_step_3.wav	xoz_step_3.wav
daerd_step_5.wav	dzaek_step_5.wav	naeft_step_5.wav	tat_step_3.wav	daerd_step_5.wav	dzaek_step_5.wav
pat_step_2.wav	pos_step_2.wav	mas_step_2.wav	tsots_step_2.wav	mard_step_2.wav	xoz_step_2.wav
pat_step_6.wav	pos_step_6.wav	mas_step_6.wav	tsots_step_6.wav	mard_step_6.wav	xoz_step_6.wav
pat_step_3.wav	pos_step_3.wav	mas_step_3.wav	tsots_step_3.wav	mard_step_3.wav	xoz_step_3.wav
pat_step_5.wav	pos_step_5.wav	mas_step_5.wav	tsots_step_5.wav	mard_step_5.wav	xoz_step_5.wav
hangach.wav	haeshk.wav	xing.wav	xats.wav	daergaex.wav	deshaek.wav
kit.wav	kich.wav	hing.wav	hats.wav	tuz.wav	tut_neutral.wav
pat_step_7.wav	pos_step_7.wav	mas_step_7.wav	tsots_step_7.wav	mard_step_7.wav	xoz_step_7.wav
pat_step_2.wav	pos_step_2.wav	mas_step_2.wav	tsots_step_2.wav	mard_step_2.wav	xoz_step_2.wav
pat_step_5.wav	pos_step_5.wav	mas_step_5.wav	tsots_step_5.wav	mard_step_5.wav	xoz_step_5.wav
pat_step_3.wav	pos_step_3.wav	mas_step_3.wav	tsots_step_3.wav	mard_step_3.wav	xoz_step_3.wav
xing.wav	xa.wav	xars.wav	xa.wav	xars.wav	xats.wav
hing.wav	ha.wav	hars.wav	ha.wav	hars.wav	hats.wav
pat_step_2.wav	pos_step_2.wav	mas_step_2.wav	tsots_step_2.wav	mard_step_2.wav	xoz_step_2.wav
pat_step_7.wav	pos_step_7.wav	mas_step_7.wav	tsots_step_7.wav	mard_step_7.wav	xoz_step_7.wav
pat_step_4.wav	pos_step_4.wav	mas_step_4.wav	tsots_step_4.wav	mard_step_4.wav	xoz_step_4.wav
daerd_step_3.wav	dzaek_step_3.wav	naeft_step_3.wav	tat_step_5.wav	daerd_step_3.wav	dzaek_step_3.wav
pat_step_6.wav	pos_step_6.wav	mas_step_6.wav	tsots_step_6.wav	mard_step_6.wav	xoz_step_6.wav
pat_step_1.wav	pos_step_1.wav	mas_step_1.wav	tsots_step_1.wav	mard_step_1.wav	xoz_step_1.wav
pat_step_4.wav	pos_step_4.wav	mas_step_4.wav	tsots_step_4.wav	mard_step_4.wav	xoz_step_4.wav
daerd_step_4.wav	dzaek_step_4.wav	naeft_step_4.wav	tat_step_4.wav	daerd_step_4.wav	dzaek_step_4.wav
pat_step_1.wav	pos_step_1.wav	mas_step_1.wav	tsots_step_1.wav	mard_step_1.wav	xoz_step_1.wav
pat_step_6.wav	pos_step_6.wav	mas_step_6.wav	tsots_step_6.wav	mard_step_6.wav	xoz_step_6.wav
pat_step_3.wav	pos_step_3.wav	mas_step_3.wav	tsots_step_3.wav	mard_step_3.wav	xoz_step_3.wav
pat_step_5.wav	pos_step_5.wav	mas_step_5.wav	tsots_step_5.wav	mard_step_5.wav	xoz_step_5.wav
xats.wav	xars.wav	xa.wav	xing.wav	xats.wav	xars.wav
hats.wav	hars.wav	ha.wav	hing.wav	hats.wav	hars.wav
tiv.wav	tuz.wav	tut_neutral.wav	kich.wav	kit.wav	tiv.wav

Figure 5.2: Blocks of the dialect categorization task (each column is one block). Pink represents [h]/[χ] stimuli, orange represents test stimuli, green represents lexical stimuli, and yellow represents [a]/[æ] stimuli. Note that the order of stimuli for the word [$t^h æt^h$]/[$t^h at^h$] is opposite from the rest of the words representing the [a]/[æ] continuum because the SEA pronunciation [$t^h at^h$] (rather than [$t^h æt^h$]) served as the basis for this continuum (see Section 4.3.2.1), such that the step numbers for this word are opposite from the other [a]/[æ] words.

The task instructions, which were displayed before both the training and test portions, are shown below in **Figure 5.3**. The response screen is shown in **Figure 5.4**. As noted in Section 3.5.2, instructions were also read to participants before each section of the experiment (see Appendix D). However, after the first participant completed the experiment, it became apparent that the instructions as written were not quite sufficient in clearly conveying how participants should respond in the dialect categorization task. Specifically, as revealed in discussions after this participant had completed the experiment, she had thought that she should choose "Gavaṛ" when the given pronunciation was used in the city of Gavaṛ, but not also in the surrounding villages, while the intended interpretation of the instructions was that Gavaṛ and its surrounding villages should be compared with other cities/regions of Armenia. She also had not fully understood which option should be chosen for variants that are used both in Gavaṛ and in other regions. Because of this misunderstanding, clarifications were added to the verbal instructions for the dialect categorization task for all remaining participants. Participants were instructed that they should choose "Gavaṛ" for variants that are characteristic of the Gavaṛ dialect (as spoken in the city and/or the surrounding villages), while they should choose "not Gavaṛ" for variants that are also used in other cities/regions of Armenia. In order to re-emphasize the latter portion of these instructions, an example was also verbally provided using the [h]/[χ] control stimuli word [hing] ('five'). Participants were told that if they think that the pronunciation [hing] can be used, for example, both in Gavaṛ and in Yerevan, then they should choose "not Gavaṛ."

Հիմա դուք մեկ առ մեկ լսելու եք որոշ բառեր: Հաշվի առնելով տվյալ բառի ձևը, արտասանությունը և հնչողությունը՝ դուք պետք է որոշեք, թե այդ բառի ձևը / արտասանությունը / հնչողությունը Գավառի բարբառին հատկանշակա՞ն է, թե՞ ոչ:

Եթե կարծում եք, որ հատկանշական է Գավառի բարբառին, սեղմեք 1:

Եթե կարծում եք, որ հատկանշական չէ Գավառի բարբառին (այսինքն՝ այդպես կարող են ասել ոչ միայն գավառցիները, այլև Հայաստանի այլ բնակիչները), սեղմեք 0:

Բոլոր բառերը նույն մարդն է արտասանելու: Դրա համար որոշումը պիտի կայացնեք ոչ թե մարդու ձայնի հիման վրա, այլ բառի ձևի / արտասանության / հնչողության հիման վրա:

Շարունակելու համար սեղմեք բացատի ստեղնը:

Figure 5.3: Instructions displayed at the beginning of the training and test portions of the dialect categorization task. The text reads, "Now you will hear some words one at a time. Considering the given word's form, pronunciation, and the way it sounds, you must decide if that word's form / pronunciation / the way it sounds is characteristic of Gavaṛ's dialect or not. If you think that it is characteristic of Gavaṛ's dialect, press 1. If you think that it is not characteristic of Gavaṛ's dialect (that is, not only people from Gavaṛ, but also other residents of Armenia can speak that way), press 0. The same person will pronounce all of the words. So, you should make your decision not based on the person's voice, but rather based on the word's form / pronunciation / the way it sounds. Press the spacebar to continue."



Figure 5.4: Response screen from the dialect categorization task. Image from MediaHub (2024).

5.3.4 Participants and data collection

The dialect categorization task was completed by the same 51 participants who completed the production task and AX task (see Section 2.4.1 for detailed information about the participants). Information about the data collection procedures for both perception tasks can be found in Section 3.5.2. As noted in Section 3.5.2, one participant's responses to both perception tasks were lost.

5.4 Overview of analyses

5.4.1 Criteria for exclusion of participants

Analyses that involved only the data from the dialect categorization task (i.e., analyses that involved no metrics of performance on the AX task) included the responses of all

participants. Though Participant 1 reported having misunderstood some aspects of the task instructions (see Section 5.3.2.2 above), it was nevertheless decided to include her responses in the analysis. This decision was reached based on comparison of her responses with those of other participants, which suggested that there may have been a subset of participants who interpreted the instructions in a similar way, and/or that Participant 1's response patterns may have been based not only on her interpretation of the instructions, but also on aspects of the stimuli that influenced some other participants' responses in a similarly systematic manner (see Section 5.6.2.1 below for a full discussion of this issue).

The potential for some other participants to have misunderstood the instructions was revealed by responses to the control conditions. There were numerous participants who did not respond in the expected way to one of the control conditions – specifically, many participants consistently answered "Gavaṛ" for both dialect and neutral stimuli for either the lexical condition or the [h]/[χ] condition. There were also a few participants who answered in this way for both control conditions. Nevertheless, participants' post-experiment feedback about the control conditions revealed that they may not have been a fully accurate gauge of participants' understanding of the task. Some participants reported that the [h]/[χ] consonants were sometimes difficult to hear, and that the intended neutral lexical items were nevertheless being pronounced with a "Gavaṛ accent" by the talker. Given these circumstances, and given that even those participants who responded in an unexpected way to both control conditions showed meaningful-looking responses to the test conditions, it was decided to include all participants in the analysis. In general, the dialect categorization task was expected to be more difficult than the AX task, as it relied on participants' metalinguistic knowledge and subjective impressions about the stimuli. Because the goal of the dialect categorization task was precisely to explore such knowledge and impressions, it cannot be said to have unambiguous correct and incorrect answers in the same way that the AX task does. It was thus decided that using a clear criterion to exclude participants (such as percent of correct responses on control stimuli) would be less appropriate.

5.4.2 Summary of analyses and predictions

The following sections present the methods and results of two sets of analyses. Section 5.5 presents group-level mixed-effects logistic regression models, which assess participants' responses to different continua steps for the various stimuli words, while also investigating the potential role of demographic factors and production metrics in predicting responses to different

continua steps. While these models are informative regarding the group-level response pattern to different continua steps and words, a different analysis method was needed in order to investigate individual-level response patterns (though random slopes were estimated in the group-level models, they were deemed unsuitable for use in a true individual-level analysis – see Section 5.6.1).

Section 5.6 describes this individual-level analysis, which began with the use of individual logistic regression models for the calculation of several metrics that characterized each participant's response pattern for each word. Based on these metrics and on visual examination of individual-level responses, participants were also sorted into different categories for each word, based on whether they showed the expected response pattern, an opposite pattern, or no pattern. The metrics derived from the individual logistic regression models, along with the categorical groupings of participants, then became dependent variables in additional mixed-effects linear and logistic regression models. These models explored differences across the two vowel pairs and various stimuli words, assessing the degree to which responses to the different vowel pairs and words matched expectations. The models also investigated whether demographic factors, production metrics, and metrics of performance on the AX task predicted the various individual-level metrics.

As in the models used to analyze the AX task (see Chapter 4), all sets of models in both portions of the analyses included predictors like STEP, VOWEL PAIR, and/or WORD, as these predictors were crucial for the core goal of the analysis – namely, the investigation of differences in how participants categorized the stimuli of the different vowel pairs and words. In contrast, predictors related to demographic factors, production metrics, and metrics of performance on the AX task are not included in the presented models in cases where they were clearly not related to responses, as these predictors can be considered the most exploratory aspect of the analyses presented in this chapter. Similarly to Chapter 4, this chapter includes discussion of cases where predictors approached but did not reach the significance threshold of $p<0.05$, in connection with the exploratory nature of the analyses and with the apparent tendency (which emerged from the analysis of both perception tasks) for the significance of predictors to differ depending on the methodological decisions made in a given analysis. Though such sub-significant predictors are identified throughout the analyses, the corresponding trends in the data are viewed as potentially

more meaningful when multiple analyses exhibit the same trend (see Section 5.7 for overview of the trends found in different analyses).

In general, it was expected that responses to the task would differ for the two vowel pairs. In particular, in connection with the social salience of the [a]/[ɒ]$_A$ pair, it was predicted that participants would respond to this vowel pair in a more expected manner. This could include both clearer categorization of the continuum endpoints (i.e., consistently responding "Gavaṛ" for the dialect-like endpoint and "not Gavaṛ" for the SEA-like endpoint) and more categorical (i.e., less noisy) overall response patterns that have a relatively clear crossover from "Gavaṛ" to "not Gavaṛ" responses. It was also expected that the crossover points from "Gavaṛ" to "not Gavaṛ" responses (as quantified by perceptual boundary location – see Section 5.6.1) could differ for the two vowel pairs, perhaps in a way that related to accuracy at different points along the continua in the AX task (e.g., perhaps being located in areas of higher AX accuracy). As noted above, analysis of potential perception-production relationships and of potential relationships between performance on the two tasks was largely exploratory, such that there were generally no specific predictions about these relationships. The same can be said about the analysis of demographic predictors, though based on the analysis of the AX task, in which people without higher education and older people generally had more accurate discrimination, it is possible that these same demographic groups will be found to have performed in a more expected manner on the dialect categorization task.

5.5 Analysis and results – Group-level logistic regression models

5.5.1 Logistic regression models – predicting responses to different steps

Responses to different continua steps were analyzed using mixed-effects logistic regression models. Like the analogous models for the AX task, these models were separated by vowel pair. The binary dependent variable was participants' responses of "Gavaṛ" vs. "not Gavaṛ." The predictors were as follows:

- **STEP –** This was a fixed predictor, coded as a continuous numeric predictor. The goal of including this predictor was to explore how participants responded to different continuum steps (1 through 7).
- **WORD –** This was a fixed categorical predictor, in which the levels were the three words representing each vowel pair. Model results in relation to this predictor indicate how

participants' overall responses ("Gavaṛ" vs. "not Gavaṛ") were different for the different words.

- **Interaction between WORD and STEP –** This interaction tells us whether the pattern of responses across different steps is different for the three different words.
- **Random intercepts for trial order within block –** This random effect accounts for the influence of the order of the presentation of the trials within each block, by accounting for the possibility that stimuli presented earlier vs. later in the block could have had different baseline responses based on their position in the block.
- **Random intercepts for trial order within task –** This random effect accounts for the influence of the overall order of the presentation of the trials on the results, allowing the model to account for the possibility that different trials could have different baseline responses based on their order in the task. As the task was blocked by word, with the blocks presented in the same order for all participants, this random effect also accounts for potential effects of the order of the presentation of the blocks.
- **Random intercepts for participants –** This random effect accounts for the possibility that different participants could have different baseline levels of responding "Gavaṛ" vs. "not Gavaṛ."
- **Random slopes for participants, relative to STEP, WORD, and their interaction –** These random slopes account for the fact that the effects of STEP, WORD, and their interaction on responses could be different for different participants.
- **Demographic factors and production metrics –** The same demographic factors and production metrics tested in the AX models were tested in these models as fixed predictors.

As with the analogous models conducted in the analysis of the AX task, the models were implemented in R (R Core Team 2021) using the *glmer()* function from the *lme4* package, while p-values were calculated using the *lmerTest* package (Laplace Approximation method). Continuous predictors were scaled. Collinearity of predictors was checked using the *vif()* function from the *car* package (Fox & Weisberg 2019), and models were checked for convergence issues using the *allFit()* function from the *lme4* package. The *emmeans(), emtrends()* and *pairs()* functions from the *emmeans* package (Lenth & Piaskowski 2026) were

used to compute and compare responses to different levels of WORD and the slopes of different levels of WORD relative to STEP. The procedure for determining which demographic and production predictors to keep in the models was the same as described for the AX models in Chapter 4. Models with and without correlations between the random effects in relation to participants were compared using ANOVA, and in all cases these correlations improved the model fit. All models presented in Section 5.5.2 thus include such correlations. As with the AX models, AIC and BIC were used as a general guide for comparing models, and discrepancies between their preferences will be discussed when relevant in Section 5.5.2 below. Differences between results in relation to production metrics calculated with Bark-converted vs. non-Bark-converted data are also discussed when relevant (specifically for cases involving ED-median).

5.5.2 Results of logistic regression models

5.5.2.1 [a]/[ɒ]$_A$

In the mixed-effects logistic regression models predicting responses to different steps of the [a]/[ɒ]$_A$ continuum, ED-median was found to be a significant predictor, while EDUCATION LEVEL was sub-significant. The directions of the effects were such that participants with larger distances between their [a] and [ɒ]$_A$ vowels in production (i.e., more dialect-like productions) chose "Gavaṛ" less, and participants without higher education chose "Gavaṛ" more. The results of the model including these predictors (as well as fixed effects of STEP, WORD, and their interaction, random intercepts for trial order within blocks, random intercepts for overall trial order, random intercepts for participants, and random slopes for participants relative to STEP, WORD, and their interaction) are presented in **Table 5.9**. The table presents a model in which ED-median was calculated from non-Bark-converted data, but the results of the model with this predictor calculated from Bark-converted data were very similar. The negative slope of STEP indicates that for the reference level *mard,*[29] "Gavaṛ" responses decreased toward the right end of the continuum (step 7, i.e. the most SEA-like step). The pairwise comparison of levels of WORD in **Table 5.10** shows that overall responses were significantly different between *pat* and *mard*, but not between any other pairs. The pairwise comparisons of the slope of STEP in relation to

[29] The reference level of WORD in all models was chosen arbitrarily, as pairwise comparisons (implemented via *emmeans()*) would nevertheless be necessary to compare all words with all other words regardless of the coding scheme chosen for the models.

each word presented in **Table 5.11** show that the slope in relation to *pat* was significantly different from the other two slopes, which were not significantly different from each other.

Table 5.9: Results of mixed-effects logistic regression model predicting responses to the dialect categorization task for [a]/[ɒ]$_A$. The model structure was as follows: **RESPONSE ~ STEP * WORD + EDUCATION LEVEL + ED-median + (STEP * WORD | participant) + (1 | order within block) + (1 | overall order)**. The reference level for WORD was *mard.* P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places.

Predictor	**estimate**	**p-value**
STEP	-0.692	*0.014*
WORD (*mas)*	-0.392	0.184
WORD (*pat)*	-0.617	*0.012*
EDUCATION LEVEL – non-higher vs. higher	0.941	0.055
ED-median	-0.569	*0.020*
STEP – WORD (*mas*) interaction	0.167	0.461
STEP – WORD (*pat*) interaction	-0.753	*0.004*

Table 5.10: Results of pairwise comparisons of the levels of WORD from the model presented in **Table 5.9**, computed using *emmeans()*. P-values have been adjusted using the Tukey method. P-values of less than 0.05 are italicized.

Pair	**p-value**
mard – mas	0.379
mard – pat	0.031
mas – pat	0.640

Table 5.11: Results of pairwise comparisons of the slope of STEP across levels of WORD from the model presented in **Table 5.9**, computed using *emtrends()*. P-values have been adjusted using the Tukey method. P-values of less than 0.05 are italicized.

Pair	**p-value**
mard – mas	0.741
mard – pat	*0.011*
mas – pat	*0.002*

Differences in the slopes of each word can be seen in the visualization of pooled data in **Figure 5.5**, which shows that *pat* has a steeper negative slope than the other two words (though all three words have a negative slope). This and all other data visualizations in this chapter were created with *ggplot2* (Wickham 2016). This figure demonstrates that at the group level, the expected response pattern (more "Gavaṛ" responses toward the left end of the continuum, i.e., step 1) was exhibited for all three words.

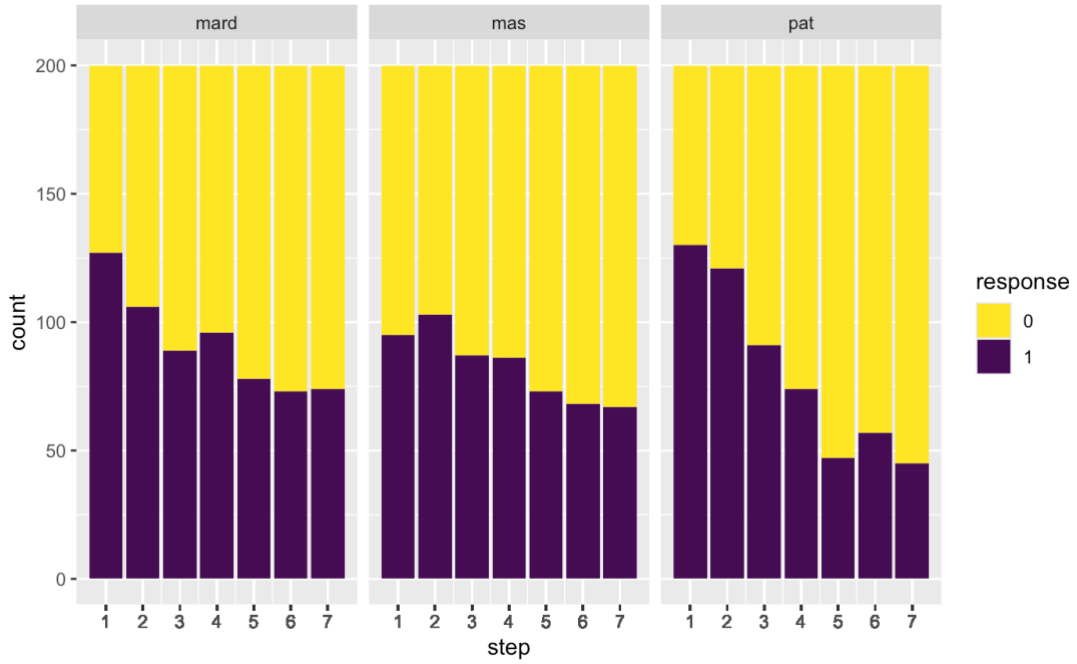


Figure 5.5: Pooled responses by WORD for [a]/[ɒ]$_A$ in the dialect categorization task. Purple (1) represents "Gavaṛ" responses, while yellow (0) represents "not Gavaṛ" responses.

Individual slopes and confidence intervals for each word calculated with *emtrends()* are presented in **Table 5.12**. The confidence intervals for *mard* and *pat* do not contain 0, suggesting that these negative slopes are significantly different from 0, while the slope for *mas* just barely contains 0 (meaning the significance of this slope is more questionable). **Figures 5.6-5.7** illustrate the trends found by the model in relation to ED-median and EDUCATION LEVEL.

Table 5.12: Slopes of STEP across levels of WORD from the model presented in Table 5.9, computed using *emtrends()*. 0.95 confidence intervals calculated with *emtrends()* are provided.

Word	**Slope**	**Lower CI**	**Upper CI**
mard	-0.692	-1.25	-0.139

mas	-0.525	-1.09	0.041
pat	-1.445	-1.99	-0.904

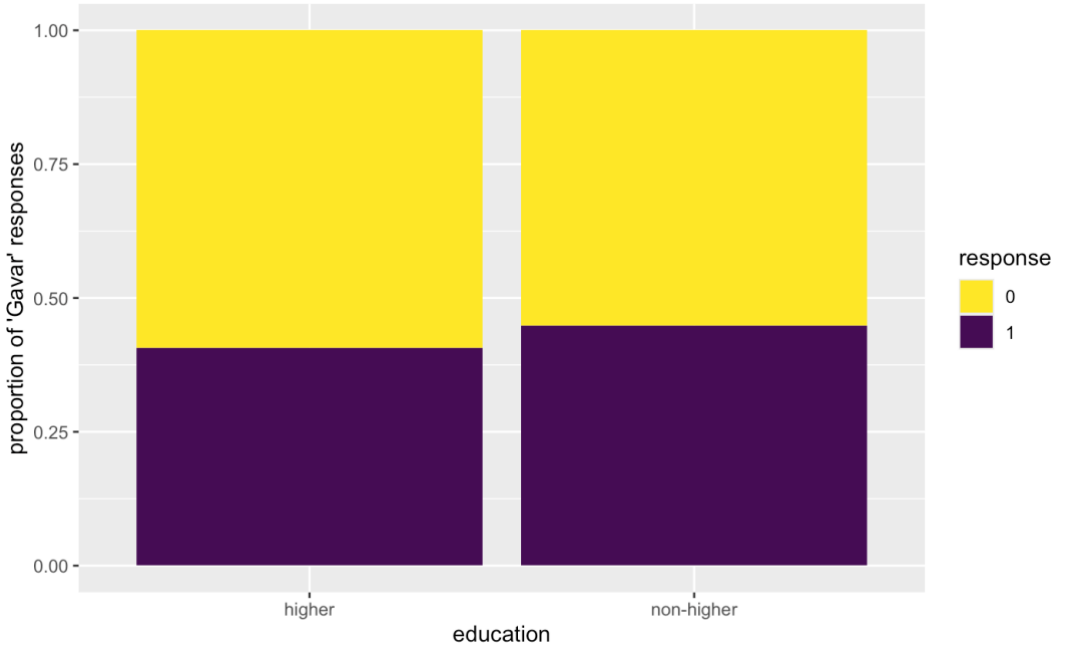


Figure 5.6: Proportion of "Gavaṛ" responses by EDUCATION LEVEL for [a]/[ɒ]$_A$ in the dialect categorization task. Purple represents "Gavaṛ" (1) responses, while yellow represents "not Gavaṛ" (0) responses. "Gavaṛ" responses made up a larger portion of the responses for the non-higher group.

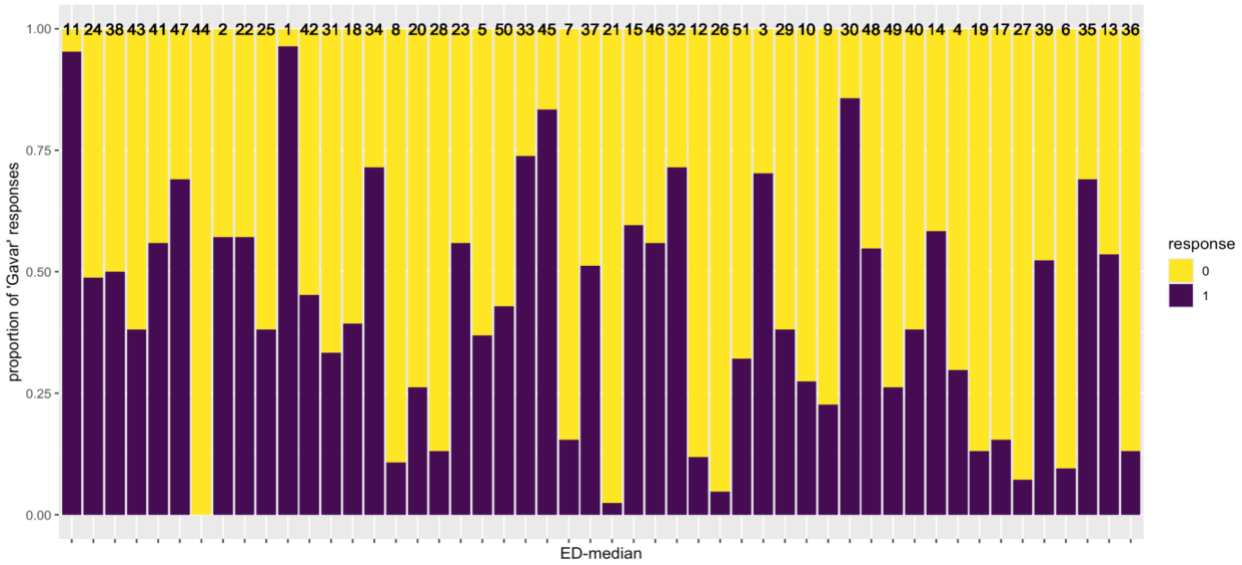


Figure 5.7: Proportion of "Gavaṛ" responses in relation to ED-median for [a]/[ɒ]$_A$ in the dialect categorization task. Purple represents "Gavaṛ" (1) responses, while yellow represents "not Gavaṛ" (0) responses. Each bar represents one participant, and the bars are arranged from smallest ED-median in production for the given vowel pair (left) to

largest ED-median (right). Participant numbers are displayed at the top of the graph. There is a weak visual trend according to which the proportion of "Gavaṛ" responses decreases as ED-median increases.

5.5.2.2 [ɔ]/[ɒ]o

In the mixed-effects logistic regression models predicting responses to different steps of the [ɔ]/[ɒ]O continuum, Pillai scores and GENDER were found to be sub-significant predictors. Participants with less overlap (i.e., larger Pillai scores) between their [ɔ] and [ɒ]O vowels in production (i.e., more dialect-like productions) tended to choose "Gavaṛ" more, as did women. The results of the model including these predictors (as well as fixed effects of STEP, WORD, and their interaction, random intercepts for trial order within blocks, random intercepts for overall trial order, random intercepts for participants, and random slopes for participants relative to STEP, WORD, and their interaction) are presented in **Table 5.13**. (Note that the model was run with Pillai scores calculated from Bark-converted data, though as discussed in Chapter 2, Bark-conversion resulted in only very small changes to Pillai scores likely caused by rounding). The negative slope of STEP indicates that for the reference level *pos,* "Gavaṛ" responses decreased toward the right end of the continuum (step 7, i.e., the most SEA-like step). The pairwise comparison of words in **Table 5.14** shows that overall responses were significantly different between *tsots* and the other two words. The pairwise comparisons of the slope of STEP in relation to each word presented in **Table 5.15** show that the slopes of the three words were not significantly different from one another.

Table 5.13: Results of mixed-effects logistic regression model predicting responses to the dialect categorization task for [ɔ]/[ɒ]O. The model structure was as follows: **RESPONSE ~ STEP * WORD + GENDER + Pillai scores + (STEP * WORD | participant) + (1 | order within block) + (1 | overall order)**. The reference level for WORD was *pos.* P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places.

Predictor	**estimate**	**p-value**
STEP	-0.559	*0.047*
WORD (*tsots)*	1.749	*<0.001*
WORD (*xoz)*	-0.392	0.203
GENDER – men vs. women	-0.777	0.074
Pillai scores	0.412	0.059

STEP – WORD (*tsots*) interaction	-0.032	0.897
STEP – WORD (*xoz*) interaction	0.382	0.057

Table 5.14: Results of pairwise comparisons of the levels of WORD from the model presented in **Table 5.13**, computed using *emmeans()*. P-values have been adjusted using the Tukey method. P-values of less than 0.05 are italicized.

Pair	**p-value**
pos – tsots	*<0.001*
pos – xoz	0.410
tsots – xoz	*<0.001*

Table 5.15: Results of pairwise comparisons of the slope of STEP across levels of WORD from the model presented in **Table 5.13**, computed using *emtrends()*. P-values have been adjusted using the Tukey method. P-values of less than 0.05 are italicized.

Pair	**p-value**
pos – tsots	0.991
pos – xoz	0.137
tsots – xoz	0.066

Figure 5.8 shows the pooled data in relation to each word, demonstrating that all three words showed the expected pattern (more "Gavaṛ" responses toward the left end of the continuum, i.e., step 1) at the group level. Nevertheless, the figure also shows a strange pattern in which the proportion of "Gavaṛ" responses at step 7 (right end of the continuum) is higher than at steps 6, 5, and sometimes 4. The implications of this pattern will be further discussed in Sections 5.5.2.3 and 5.7.1.

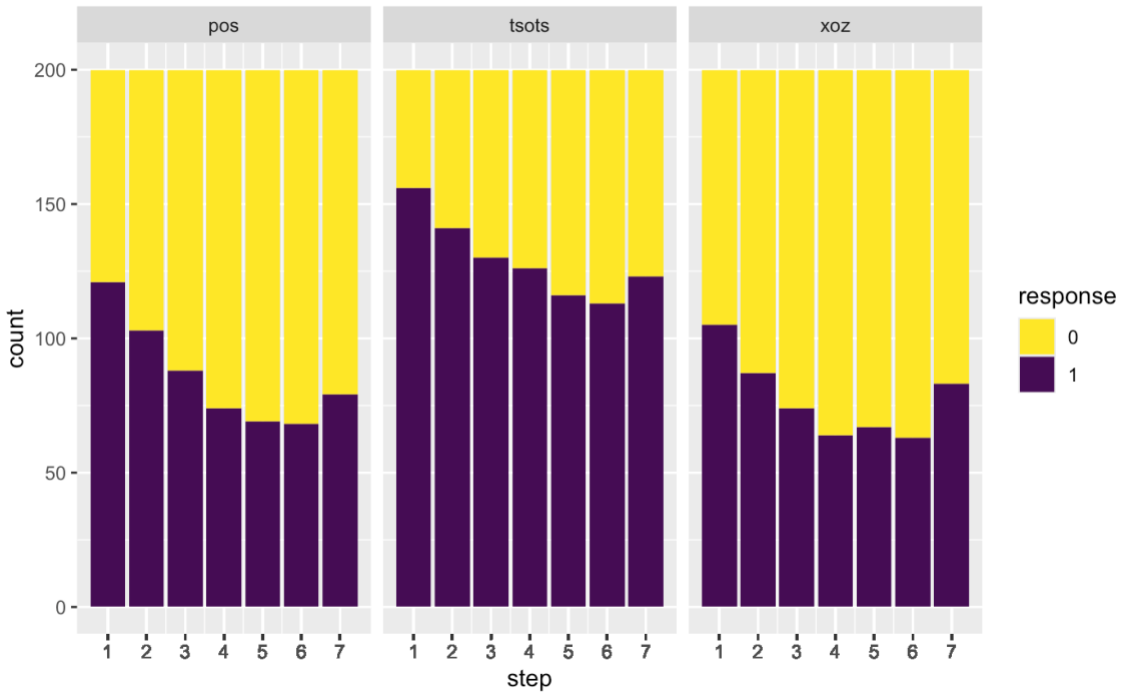


Figure 5.8: Pooled responses by WORD for [ɔ]/[ɒ]$_O$ in the dialect categorization task. Purple (1) represents "Gavaṛ" responses, while yellow (0) represents "not Gavaṛ" responses.

Individual slopes and confidence intervals for each word calculated with *emtrends()* are presented in **Table 5.16**. The confidence intervals for *pos* and *tsots* do not contain 0, suggesting that these negative slopes are significantly different from 0 (although the upper confidence interval for *pos* is extremely close to zero). On the other hand, the upper confidence interval of *xoz* is well above zero. **Figures 5.9-5.10** illustrate the sub-significant trends found by the model in relation to GENDER and Pillai scores.

Table 5.16: Slopes of STEP across levels of WORD from the model presented in **Table 5.13**, computed using *emtrends()*. 0.95 confidence intervals calculated with *emtrends()* are provided.

Word	**Slope**	**Lower CI**	**Upper CI**
pos	-0.559	-1.110	-0.008
tsots	-0.591	-0.917	-0.265
xoz	-0.177	-0.621	0.267

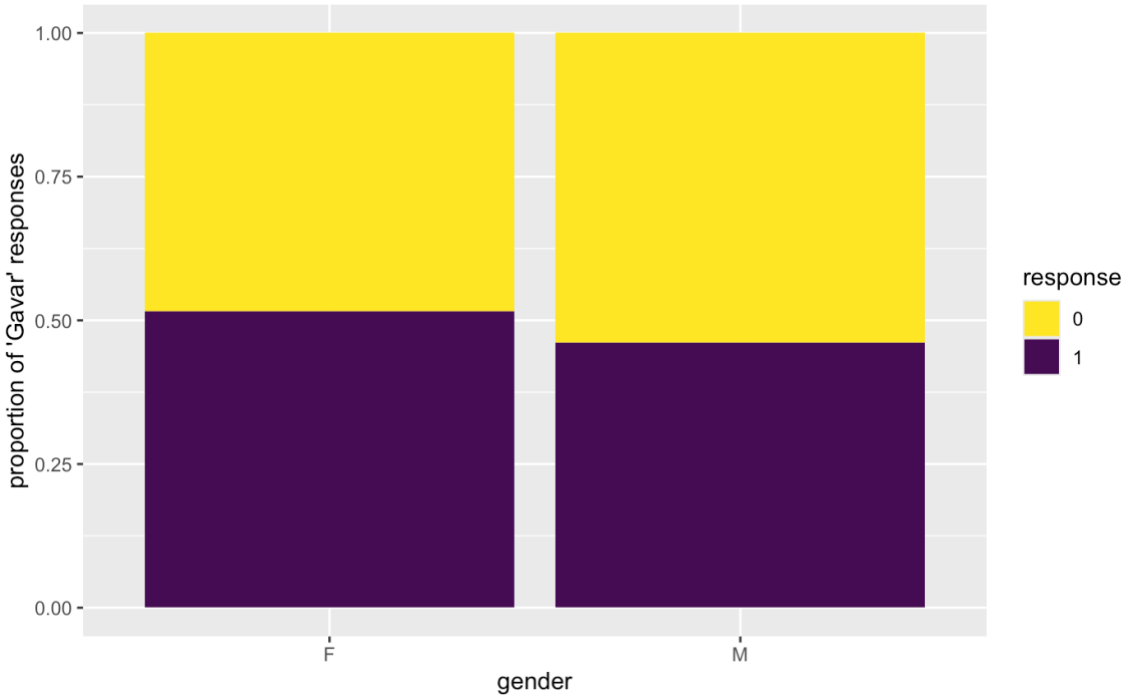


Figure 5.9: Proportion of "Gavaṛ" responses by GENDER for [ɔ]/[ɒ]$_O$ in the dialect categorization task. Purple represents "Gavaṛ" (1) responses, while yellow represents "not Gavaṛ" (0) responses. "Gavaṛ" responses made up a slightly larger portion of the responses for women in comparison with men.

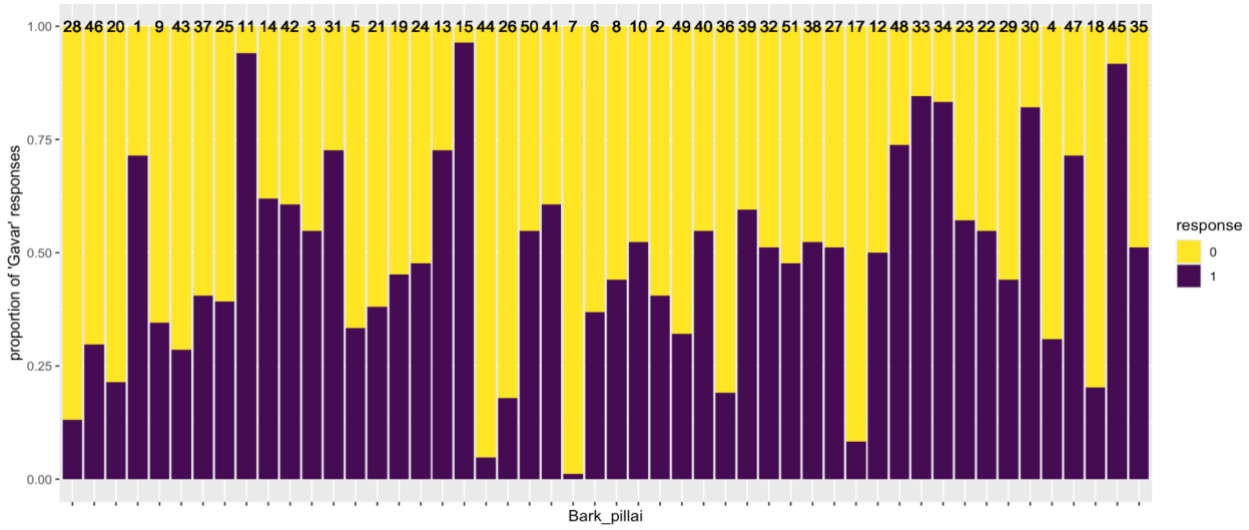


Figure 5.10: Proportion of "Gavaṛ" responses in relation to Pillai scores for [ɔ]/[ɒ]$_O$ in the dialect categorization task. Purple represents "Gavaṛ" (1) responses, while red represents "not Gavaṛ" (0) responses. Each bar represents one participant, and the bars are arranged from smallest Pillai score in production for the given vowel pair (left) to largest Pillai score (right). Participant numbers are displayed at the top of the graph. According to the model presented in Table 5.13, there is a weak trend according to which proportion of "Gavaṛ" responses increases as Pillai scores increases (i.e., as overlap between vowel categories in production decreases).

5.5.2.3 Summary of results of group-level logistic regression models

As discussed in Sections 5.5.2.1 and 5.5.2.2 above, the logistic regression models predicting responses to different continuum steps sometimes found differences between the overall responses for different words within each vowel pair, and differences between the slopes of STEP in relation to each word. For example, the overall responses and STEP slope for the word *pat* were significantly different from some of the other words representing [a]/[ɒ]$_A$, and the overall responses for *tsots* were significantly different from the other two words representing [ɔ]/[ɒ]$_O$.

All three words representing each vowel pair had a negative slope according to the models, such that group-level responses for all six stimuli words aligned with the expected response pattern of having more "Gavaṛ" responses toward the dialect-like end of the continuum (step 1). Nevertheless, visualizations of the pooled data (**Figures 5.5** and **5.8**) suggested that the responses for [ɔ]/[ɒ]$_O$ were noisier, as "Gavaṛ" responses increased slightly toward the SEA-like end of the continuum (step 7), rather than steadily decreasing across the continuum steps. Examination of the estimated slopes and their confidence intervals also suggested that the slopes for [a]/[ɒ]$_A$ were overall slightly more significantly different from 0 than those for [ɔ]/[ɒ]$_O$ (see the general discussion in Section 5.7.1 for elaboration of this point). Thus, the results of the logistic regression models and pooled visualizations provide support for the hypothesis that the social salience of [a]/[ɒ]$_A$ would lead to it being categorized in a more expected manner. The evidence provided by these models is nevertheless very tenuous, especially given the vast individual variation in participants' responses to this task (see further discussion in Section 5.6 below). Therefore, in order to further explore the nuances of how responses to the different vowel pairs differed, additional analyses of individual-level metrics of performance on the task were undertaken, and are presented in Section 5.6.

The logistic regression models also found some weak trends in relation to demographic factors and production metrics, though these trends are difficult to interpret given that these models simply predicted the response to each continuum step, rather than some metric of overall performance on the task. For example, in the models predicting responses to [a]/[ɒ]$_A$, participants without higher education chose "Gavaṛ" more overall, while in the models predicting responses to [ɔ]/[ɒ]$_O$, women chose "Gavaṛ" more. Both trends were very weak (sub-significant), and it is also difficult to come up with explanations as to why the given demographic groups would

exhibit the given response patterns for only the given vowel pairs. The trends found in relation to production metrics were opposite for the two vowel pairs – for [a]/[ɒ]$_A$, participants with greater distances between the relevant vowel categories in production (i.e., more dialect-like pronunciations) chose "Gavaṛ" less (significant trend), while for [ɔ]/[ɒ]$_O$, those with less overlap in production (i.e., more dialect-like pronunciations) chose "Gavaṛ" more (sub-significant). Again, because of the structure of the given models, these trends reflect overall responses to the entire continuum, rather than responses to particular continuum steps or the shapes of response patterns. It is thus unclear to what extent the given trends should be considered to contribute meaningfully to the interpretation of participants' performance on the task, especially since there is no clear explanation for why the trends would be opposite for the two vowel pairs.

5.6 Analysis and results – characterizing and predicting individual-level metrics and response patterns

5.6.1 Individual logistic regression models – calculating slopes and perceptual boundary locations

As noted in the previous section, the large logistic regression models predicting responses on the dialect categorization task included random slopes for participants in relation to WORD, STEP, and their interaction. It was hoped that these random slopes could be used for an individual-level analysis of performance on the task – for example, with the goal of comparing how many participants had the expected vs. opposite response pattern for different words. However, extraction and examination of these random slopes showed that for many individual participants, the random slopes deviated considerably from visual patterns observed in the raw data. For example, a participant who had the expected response pattern (negative slope) for one word, the opposite pattern (positive slope) for another word, and no pattern (flat slope) for the third word representing a given vowel pair would have strongly negative random slopes for all three words in the model. This likely occurred due to the high variability in response pattern both within and across participants. As Korner-Nievergelt et al. (2015: 112-133 [Chapter 7]) explain, random effects for groups (i.e., in the case of the present study, individual participants) who have highly variable data are shrunk toward the population-level effect. This is because mixed-effects models use a method called "partial pooling" in the estimation of random effects to account for both population-level and group-level tendencies and avoid overestimation of the variance

between groups. Thus, while the mixed-effects models described above are useful for modelling population-level behavior, the variability of the data set makes it such that the random slopes are not particularly useful for understanding how individual participants actually responded to the stimuli for different words.

In light of these issues, an alternative method for understanding individual-level responses was sought. Butcher (2022) analyzed individual-level responses in a similar perception task using logistic regression models that were run individually for each participant for each word. In Butcher's case, the experimental task also involved categorization of stimuli whose vowels varied on a continuum; however, the endpoints represented different words, such that the task involved phonemic categorization. Butcher (2022: 175-176) used the output of the logistic regression models to determine whether each participant could be said to have a categorical perceptual boundary, as well as to calculate the width and location of that boundary.

Due to the differing nature of Butcher's experiment (phonemic categorization) and the dialect categorization task in the present study, several modifications were made to Butcher's methods. Butcher (2022:175) defined the width of the perceptual boundary as the distance (in continuum steps) between the point along the continuum where the participant had a 75% probability of perceiving a given word and the point where they had a 25% probability of perceiving that word, and participants for whom this distance exceeded the width of the continuum were determined to not have perceived the stimuli in a categorical manner. However, as the categorization task in the present study did not involve phonemic perception, such a method was deemed inappropriate. There were some participants (e.g., the left panel of **Figure 5.11**) whose responses to the dialect categorization task appeared to exhibit the expected pattern (of having more "Gavaɻ" responses toward the dialect-like end of the continuum and more "not Gavaɻ" responses toward the SEA-like end), but in a relatively shallow manner – for example, without responding "Gavaɻ" 100% of the time for any continuum step. Because of this more shallow pattern shape, Butcher's boundary-width method would deem these participants as not having perceived the stimuli in a categorical manner. Nevertheless, because the task did not involve phonemic categorization, and because participants may have had their own baselines in terms of their overall tendencies to choose "Gavaɻ" or "not Gavaɻ," it seems that having a strict categorical pattern that goes from 100% "Gavaɻ" at one end of the continuum to 0% "Gavaɻ" at the other end should not be a requirement in order for a participant to be considered to have the

expected response pattern. Thus, rather than calculating boundary width according to Butcher's method, it was determined that the slopes of participants' responses (relative to STEP) should simply be collected from each model and examined in further analyses.

Attempting to extract these slopes from simple logistic regression models nevertheless proved problematic. Because some participants responded in an extremely categorical manner (e.g., right panel of **Figure 5.11** below), some models faced the issue of "complete separation" (see Suhas et al. 2023), which led to very large parameter estimates in comparison with models that did not face this issue. This issue would have made it impossible to compare the slopes extracted from different models in any quantitative way, as the slopes from models with complete separation were extreme outliers in comparison with other slopes. In order to circumvent this issue, Firth's penalized logistic regressions (see Suhas et al. 2023) were used instead and were implemented in R using the *logistf()* function from the *logistf* package (Heinze et al. 2025). As noted above, these models were run separately for each participant for each word, with STEP (coded as a continuous numeric predictor) being the only predictor. Though perceptual boundary widths were not calculated, Butcher's (2022: 176) method for calculating boundary location was nevertheless used. BOUNDARY LOCATION was defined as the place along the continuum where the participant had a 50% probability of perceiving one vs. the other variant, and was calculated using the following formula:

(log(0.5/(1-0.5)) + model intercept)/(–coefficient)

Though this method was implemented with the goal of having a quantitative way to compare approximately where each participant crossed over from hearing dialect to non-dialect variants, its results must nevertheless be interpreted with caution in the context of the present study. As noted above, because the task did not involve phonemic categorization, it could not necessarily be expected that participants would exhibit a strictly categorical response pattern, and more shallow types of response patterns should also be considered meaningful. In light of this, the given definition of the perceptual boundary may be difficult to interpret for those participants with more shallow patterns (indeed, for such participants, the point at which they would have a 50% chance of perceiving one vs. the other variant is often at a location that is not within the bounds of the continuum [i.e., the calculated value is less than 1 or greater than 7]). Thus, while

the results of the analysis of BOUNDARY LOCATION will be presented and discussed, it is important to keep in mind that it does not capture all the nuances of the variation in participants' response patterns.

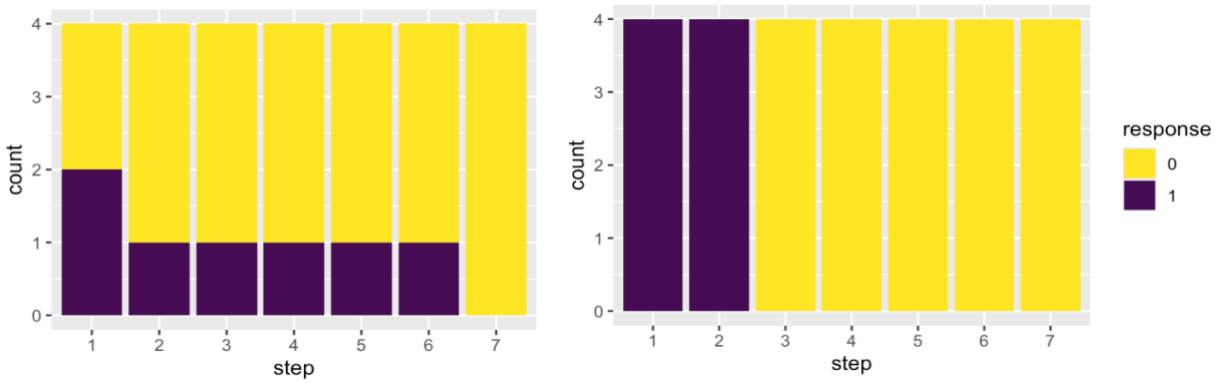


Figure 5.11: Example of shallow response pattern (left – Participant 9, *mard*) vs. "complete separation" response pattern (right – Participant 9, *mas*). Purple (1) represents "Gavaṛ" responses, while yellow (0) represents "not Gavaṛ" responses. Step 1 represents the most dialect-like variant, while step 7 represents the most SEA-like variant.

5.6.2 Categorizing participants based on response pattern

While the slopes from the individual logistic regression models described in the previous section provided a quantifiable way of comparing participants' responses, the steepness of the slopes relative to one another sometimes deviated sharply from visual impressions about different participants' response patterns. The most striking example of this issue came from instances where a participant answered "Gavaṛ" or "not Gavaṛ" for a single repetition of one of the endpoint continua steps (see **Figure 5.12** for an example). In cases such as these, the slope extracted from the logistic regression model was (-)0.65, which is relatively steep in comparison with other slopes and in comparison with the visual impression that such participants cannot be said to have shown a clearly categorical response pattern. Since the slope computed for such participants represented their response to a single repetition of a single stimulus, which could have been a mistake or unintended response, it was decided that such participants should not be considered to have demonstrated the expected (or opposite) response pattern, but rather should be considered similar to those participants whose computed slope was close to zero.

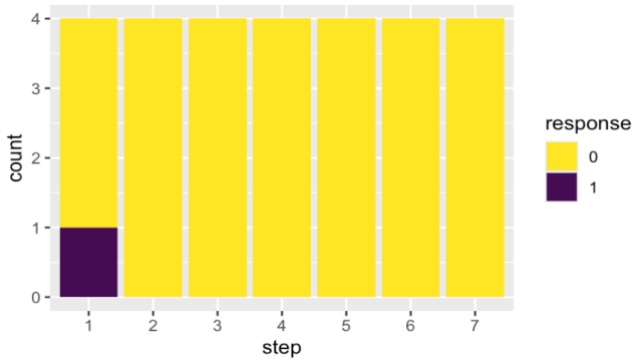


Figure 5.12: Participant 26's responses for *pat.* Purple (1) represents "Gavaṛ" responses, while yellow (0) represents "not Gavaṛ" responses. Step 1 represents the most dialect-like variant, while step 7 represents the most SEA-like variant. Though the computed slope was -0.65, the participant was deemed to have neither the expected nor opposite response pattern.

The consideration of responses such as the one shown in **Figure 5.12** raised a larger issue of how to determine which slopes were close enough to zero to be considered functionally equivalent to zero, and of how to determine what to do in other cases where the computed slope seemed incongruent with visual impressions of the responses. Likely due to factors such as the small sample size used in each model and the response shape assumed by a logistic regression model, there were additional cases where relatively large slopes were computed for response patterns that looked fairly random, while relatively small slopes were computed for patterns that visually had the expected (negative slope) or opposite (positive slope) shape, but were shallow and/or noisy. An example of the former can be seen in **Figure 5.13**, while an example of the latter is shown in **Figure 5.14**.

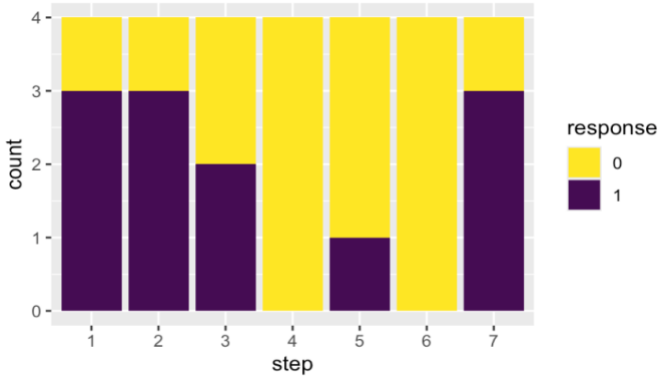


Figure 5.13: Participant 14's responses for *mard.* Purple (1) represents "Gavaṛ" responses, while yellow (0) represents "not Gavaṛ" responses. Step 1 represents the most dialect-like variant, while step 7 represents the most SEA-like variant. Though the computed slope was -0.24, the participant was deemed to have neither the expected nor opposite response pattern due to the fact that the pattern appears random, and the response to the two endpoints of the continuum was equal.

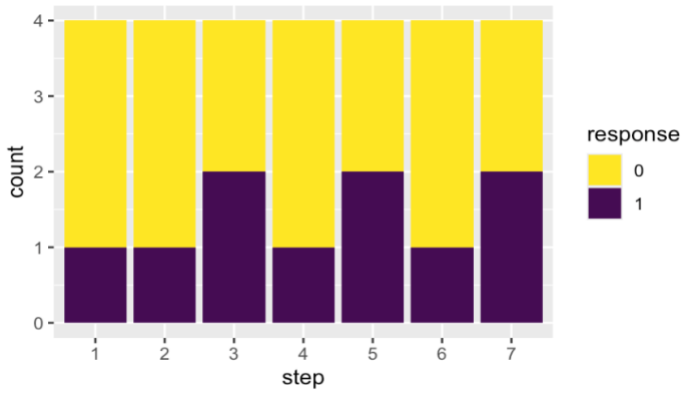
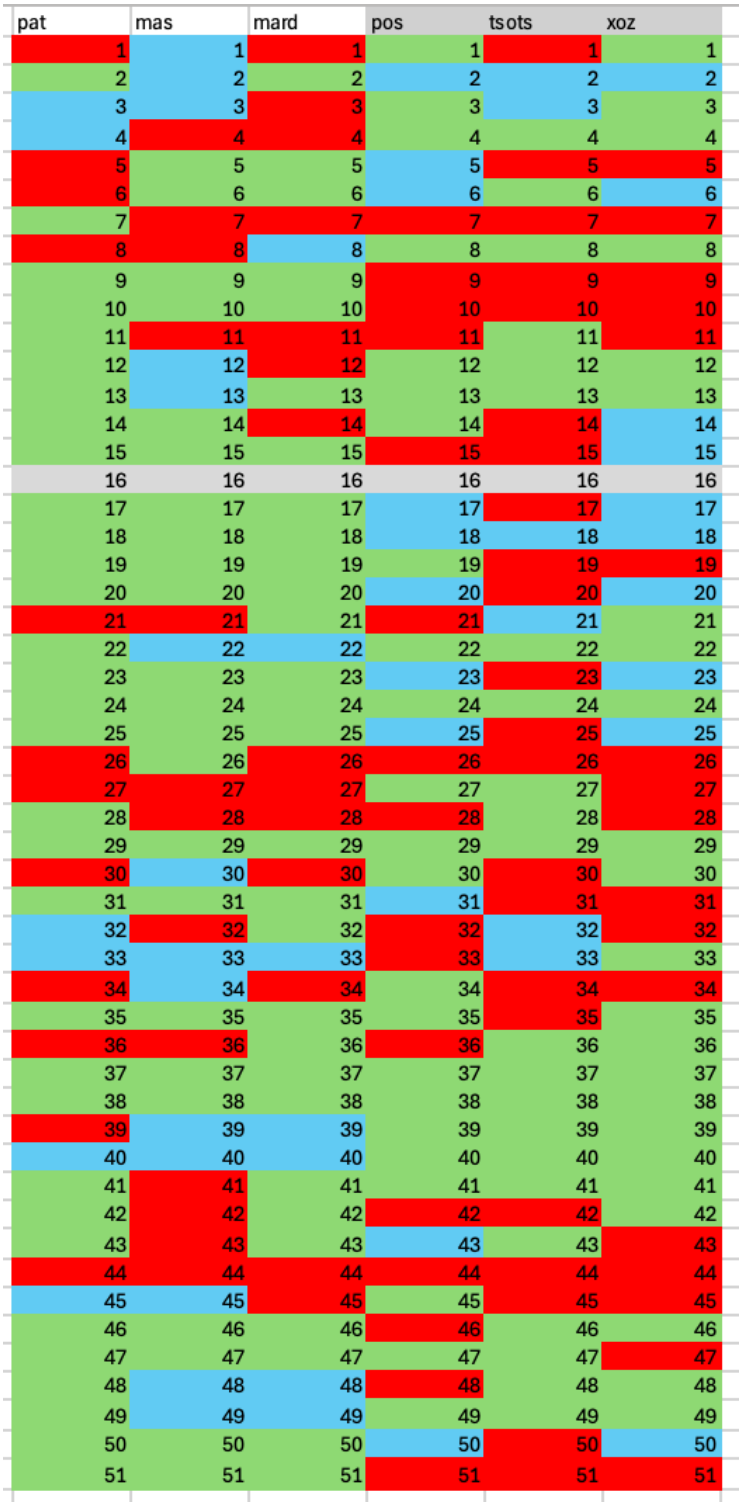


Figure 5.14: Participant 2's responses for *tsots*. Purple (1) represents "Gavaŗ" responses, while yellow (0) represents "not Gavaŗ" responses. Step 1 represents the most dialect-like variant, while step 7 represents the most SEA-like variant. Though the computed slope was 0.11, and the participant's response pattern is shallow and noisy, the participant was deemed to have the "opposite" response pattern, due to the fact that "Gavaŗ" responses were generally higher toward one end of the continuum than the other (in particular, due to the cluster of two low responses at the left end of the continuum).

Accordingly, a combination of visual examination and consideration of the computed slope values was used to sort participants into those with the expected response pattern, those with the opposite pattern, and those with no pattern. In general, those deemed to have the expected or opposite pattern had slopes with an absolute value of around 0.2 or greater, while those deemed to have no pattern had slopes whose absolute value was less than around 0.2, but exceptions were made for cases such as those depicted in **Figures 5.12-5.14** (visualizations of all participants responses can be found in Appendix E). In general, an attempt was made to be as inclusive as possible in assigning either the expected or opposite pattern – in other words, any pattern that even vaguely resembled being higher toward one end of the continuum and lower toward the other was included. Since those with very shallow and/or noisy patterns generally had very small slopes, the shallowness/noisiness of such patterns in comparison with more clear/categorical patterns would be accounted for in the quantitative comparison of slopes described in Section 5.6.3. While the method of incorporating visual impressions in order to sort participants by response pattern is undoubtedly somewhat subjective, the sorting was checked multiple times to ensure consistency across the different words. The overall results of this analysis can be seen in **Figure 5.15** (presentation of each participant's response pattern for each word, in order of participation) and **Figure 5.16** (presentation of the same information arranged by type and sub-type of response pattern). As **Figure 5.16** shows, the single most expected response type was having the expected response pattern for all three [a]/[ɒ]$_A$ words and none of the [ɔ]/[ɒ]$_O$ words (N=11), but a wide variety of response patterns was nonetheless found.



pat	mas	mard	pos	tsots	xoz
1	1	1	1	1	1
2	2	2	2	2	2
3	3	3	3	3	3
4	4	4	4	4	4
5	5	5	5	5	5
6	6	6	6	6	6
7	7	7	7	7	7
8	8	8	8	8	8
9	9	9	9	9	9
10	10	10	10	10	10
11	11	11	11	11	11
12	12	12	12	12	12
13	13	13	13	13	13
14	14	14	14	14	14
15	15	15	15	15	15
16	16	16	16	16	16
17	17	17	17	17	17
18	18	18	18	18	18
19	19	19	19	19	19
20	20	20	20	20	20
21	21	21	21	21	21
22	22	22	22	22	22
23	23	23	23	23	23
24	24	24	24	24	24
25	25	25	25	25	25
26	26	26	26	26	26
27	27	27	27	27	27
28	28	28	28	28	28
29	29	29	29	29	29
30	30	30	30	30	30
31	31	31	31	31	31
32	32	32	32	32	32
33	33	33	33	33	33
34	34	34	34	34	34
35	35	35	35	35	35
36	36	36	36	36	36
37	37	37	37	37	37
38	38	38	38	38	38
39	39	39	39	39	39
40	40	40	40	40	40
41	41	41	41	41	41
42	42	42	42	42	42
43	43	43	43	43	43
44	44	44	44	44	44
45	45	45	45	45	45
46	46	46	46	46	46
47	47	47	47	47	47
48	48	48	48	48	48
49	49	49	49	49	49
50	50	50	50	50	50
51	51	51	51	51	51

Figure 5.15: All participants' response patterns by word. Green represents the expected pattern, blue represents the opposite pattern, and red represents no pattern (Participant 16 is the participant whose responses to the experiment were lost). The left three columns are words representing [a]/[ɒ]$_{A}$, while the right three columns represent [ɔ]/[ɒ]$_{O}$.

group	pat	mas	mard	pos	tsots	xoz	sub-type
all expected	24	24	24	24	24	24	n/a
	29	29	29	29	29	29	n/a
	37	37	37	37	37	37	n/a
	38	38	38	38	38	38	n/a
5 expected	35	35	35	35	35	35	[a] more expected
	46	46	46	46	46	46	[a] more expected
	47	47	47	47	47	47	[a] more expected
	13	13	13	13	13	13	[ɔ] more expected
	41	41	41	41	41	41	[ɔ] more expected
4 expected	19	19	19	19	19	19	[a] more expected
	12	12	12	12	12	12	[ɔ] more expected
	22	22	22	22	22	22	[ɔ] more expected
	49	49	49	49	49	49	[ɔ] more expected
3 expected	9	9	9	9	9	9	[a] more expected
	10	10	10	10	10	10	[a] more expected
	15	15	15	15	15	15	[a] more expected
	17	17	17	17	17	17	[a] more expected
	18	18	18	18	18	18	[a] more expected
	20	20	20	20	20	20	[a] more expected
	23	23	23	23	23	23	[a] more expected
	25	25	25	25	25	25	[a] more expected
	31	31	31	31	31	31	[a] more expected
	50	50	50	50	50	50	[a] more expected
	51	51	51	51	51	51	[a] more expected
	4	4	4	4	4	4	[ɔ] more expected
	8	8	8	8	8	8	[ɔ] more expected
	39	39	39	39	39	39	[ɔ] more expected
	40	40	40	40	40	40	[ɔ] more expected
	6	6	6	6	6	6	mixed – [a] more expected
	14	14	14	14	14	14	mixed – [a] more expected
	42	42	42	42	42	42	mixed – [a] more expected
	43	43	43	43	43	43	mixed – [a] more expected
	36	36	36	36	36	36	mixed – [ɔ] more expected
	48	48	48	48	48	48	mixed – [ɔ] more expected
2 expected	2	2	2	2	2	2	[a] more expected
	5	5	5	5	5	5	[a] more expected
	1	1	1	1	1	1	[ɔ] more expected
	3	3	3	3	3	3	[ɔ] more expected
	27	27	27	27	27	27	[ɔ] more expected
	30	30	30	30	30	30	[ɔ] more expected
	11	11	11	11	11	11	equal
	21	21	21	21	21	21	equal
	28	28	28	28	28	28	equal
1 expected	7	7	7	7	7	7	[a] more expected
	26	26	26	26	26	26	[a] more expected
	32	32	32	32	32	32	[a] more expected
	33	33	33	33	33	33	[ɔ] more expected
	34	34	34	34	34	34	[ɔ] more expected
	45	45	45	45	45	45	[ɔ] more expected
0 expected / no data	44	44	44	44	44	44	n/a
	16	16	16	16	16	16	n/a

Figure 5.16: Participants' response patterns by word, arranged by the number of words with the expected pattern and the vowel pair that had more expected response patterns for each participant. The group of participants who had the expected pattern for three words is divided into those whose three expected patterns all belonged to the same vowel pair, and those with more mixed response patterns.

5.6.2.1 Justification of inclusion of all participants in analyses

Though it could be argued that some participants should have been excluded from the analysis of this task due to possible or confirmed misunderstanding of the instructions (see Sections 5.4.1 and 5.3.3.2 above), it was ultimately decided that all participants should be included.

As noted in Section 5.3.3.2, discussions with Participant 1 after her completion of the experiment revealed that she had misunderstood some of the instructions (which were clarified for all remaining participants). Nevertheless, examination of her responses to the test stimuli revealed an unexpected pattern that was also found among some other participants – namely, the tendency to respond with the expected pattern more so for the [ɔ]/[ɒ]$_O$ words than for the [a]/[ɒ]$_A$ words. Though this pattern is very unexpected due to [a]/[ɒ]$_A$ being the more socially salient (and more acoustically distinct) vowel pair, it was found to some extent among at least 10 other participants (see **Figure 5.16** above, which also shows that the opposite pattern of showing more expected responses for [a]/[ɒ]$_A$ than for [ɔ]/[ɒ]$_O$ was more widespread in comparison). In combination with some participants' commentary that stimuli near the SEA-like end of the continuum for the word *pos* sounded like [pus] (see Section 4.7.1), Participant 1's misunderstanding of the instructions provides a potential explanation as to why a subset of the participants responded in this way. Specifically, based on what is known about her misunderstanding of the instructions (i.e., the fact that she thought she should choose "Gavaṛ" for any pronunciations used by anyone in Gavaṛ, rather than only for dialectal pronunciations), the fact that she chose "Gavaṛ" for stimuli near the dialect-like end of the [ɔ]/[ɒ]$_O$ continuum and "not Gavaṛ" for stimuli near the SEA-like end suggests that she thought the stimuli near the SEA-like end did not sound like any pronunciations used by anyone in Gavaṛ (i.e., even by people who are not viewed as conservative dialect speakers). In other words, given that we know that some participants perceived an [u] vowel in these stimuli, it is possible that she simply found the variants near the SEA-like end of the continuum to sound strange and unfamiliar, and chose "not Gavaṛ" for this reason. Among the remaining participants who exhibited a similar pattern, the lack of expected categorization of the more salient [a]/[ɒ]$_A$ pair (which was responded to in the expected manner by a higher number of participants – see Section 5.6.4.4) suggests that these participants had lower metalinguistic awareness than other participants. This makes it surprising that they had more successful categorization of the less salient [ɔ]/[ɒ]$_O$ pair. It can thus be

proposed that these participants demonstrated the expected pattern for the [ɔ]/[ɒ]ᴏ pair for a similar reason as what has been proposed for Participant 1 – namely, that they found the dialect-like variants to sound usual and familiar (thus choosing "Gavaṛ") and the SEA-like variants to sound strange and perhaps [u]-like (thus choosing "not Gavaṛ"). It is also possible that some of these participants misunderstood the instructions in a manner similar to Participant 1, though we cannot know for sure if this was the case. In light of these complexities, it was decided that none of these participants should be excluded from the analysis, as they seem to represent an important subset of participants who may have sorted variants (at least for [ɔ]/[ɒ]ᴏ) based on their general familiarity, rather than based on metalinguistic knowledge about dialect features.

5.6.3 Linear and logistic regression models – predicting slopes, response patterns, and boundary locations with AX performance, production metrics, demographics

Additional mixed-effects linear and logistic regression models were run with the goal of exploring trends in participants' individual response patterns, slopes, and boundary locations, including (a) differences between the different words and vowel pairs, (b) trends in relation to demographic factors, (c) trends related to production metrics, and (d) trends in relation to metrics of performance on the AX task. These models thus had the goal not only of exploring group and individual-level performance on the dialect categorization task across different words and vowel pairs, but also of exploring individual perception-production relationships and individual performance across the two tasks (these latter two topics are discussed in more detail in Chapter 6). While the large logistic regression models described in Section 5.5 predicted responses to different continua steps, the structure of the models described in this section removes the need for the STEP predictor, allowing the two vowel pairs to be directly compared within the same model.

Aside from the obligatory WORD and VOWEL PAIR predictors, more exploratory predictors tested in these models included the same demographic factors (GENDER, BIRTH YEAR, EDUCATION LEVEL, PREVIOUS, HATSARAT) and production metrics (Pillai scores, ED-median calculated from Bark-converted and non-Bark-converted data) tested in the linear models presented in Chapter 4, as well as metrics of performance on the AX task calculated in Chapter 4 (ACCURACY, *d'*, SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED [both numeric and in Bark]). These predictors are summarized in **Table 5.17**. Metrics that were highly collinear with one another were tested in separate models.

Table 5.17: Summary of demographic, production, and AX predictors tested in linear and logistic regression models predicting metrics of performance on the dialect categorization task.

demographic	**production**	**AX performance**
• GENDER • BIRTH YEAR • EDUCATION LEVEL • PREVIOUS • HATSARAT	• Pillai scores • ED-median (Bark-converted) • ED-median (non-Bark-converted)	• ACCURACY • *d'* • SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED (numeric) • SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED (Bark)

The inclusion of metrics from the AX task created difficulties in connection with the fact that these metrics were computed separately for each of the four stimuli words included in that task (*pat, mas, pos, tsots*), and obviously could not be computed for the words that were only included in the dialect categorization task (*mard, xoz*). Thus, the word-level metrics computed for the AX task could not be included as predictors in models predicting performance on all six stimuli words in the dialect categorization task. In order to address this issue, two sets of models were run – one predicting performance on the dialect categorization task for only those words that were also stimuli in the AX task, with word-level AX metrics as predictors, and the other predicting performance on all words in the dialect categorization task, with pooled AX metrics for each vowel pair as predictors. The goal of running both sets of models was to compare results across the two sets of stimuli and to guard against overinterpretation of the predictive power of the AX metrics, which would be a risk if we were to consider only models in which the pooled AX metrics predict dialect categorization performance for a set of words that were not all included as AX stimuli. For ACCURACY and *d'*, the pooled metrics were calculated by pooling the data of the two words representing each vowel pair in the AX task, while for SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED (both numeric and in Bark), the pooled metric was calculated by simply taking the average of the interval values for the two words representing each vowel pair.

For reasons similar to those described for the linear models in Section 4.5, both pooled models (models with both vowel pairs) and separate models (models with only one vowel pair) were also run. Thus, for each outcome variable, six sets of models were run: pooled models with only the four repeated stimuli words, pooled models with all six stimuli words, separate models (for [a]/[ɒ]$_A$ and [ɔ]/[ɒ]$_O$) with only repeated words, and separate models with all six words).

Because the results of the different sets of models were often redundant, pointing to the same general trends in the data, only the pooled models will be presented in Section 5.6.4 below. In cases where trends observed in the separate models motivated the testing of certain interactions in the pooled models, this will be noted.

Models were run with five different outcome variables – SLOPE, ABSOLUTE VALUE OF SLOPE, SLOPE only among participants with the expected response pattern, RESPONSE PATTERN (categorical response variable), and BOUNDARY LOCATION. SLOPE was examined in 3 different ways in order to assess its general values, steepness (regardless of the direction of the response pattern), and steepness only among those participants who had exhibited the expected pattern. As noted above, six sets of models were generally run for each outcome variable; however, for SLOPE only among participants with the expected response pattern, only pooled models were run, as the smaller size of the data set for this outcome variable caused singularity issues in non-pooled models. The different outcome variables, as well as a summary of which participants were included in the models predicting which variables, are presented in **Table 5.18**.

Table 5.18: Overview of outcome variables of linear and logistic regression models predicting performance on the dialect categorization task.

Outcome variable	**SLOPE**	**ABSOLUTE VALUE OF SLOPE (steepness)**	**SLOPE among those with expected pattern**	**RESPONSE PATTERN (categorical)**	**BOUNDARY LOCATION**
Participants included	all, except those whose computed relatively sleep slope was deemed to not match flat/random visual pattern	all, except those whose computed relatively sleep slope was deemed to not match flat/random visual pattern	participants with expected response pattern	all	All participants with either "expected" or "opposite" visual response pattern
Number of participants excluded on basis of incongruent slope and visual pattern	*pat* = 2; *mas* = 0; *mard* = 3; *pos* = 2; *tsots* = 4; *xoz* = 2	*pat* = 2; *mas* = 0; *mard* = 3; *pos* = 2; *tsots* = 4; *xoz* = 2	n/a	n/a	n/a

For the outcome variable of SLOPE among only those participants who had exhibited the expected pattern, only pooled (rather than separate) models were run, as the separate models had issues with singularity due to the smaller data set used in these models. The models predicting SLOPE and ABSOLUTE VALUE OF SLOPE included all participants except those such as the ones

depicted in **Figure 5.12** and **Figure 5.13** above, where the participant was deemed to have neither the expected nor opposite response pattern despite a relatively steep slope having been extracted from the individual logistic regression models. In other words, the models excluded those participants for whom the relatively high slope did not match the visual impression of a flat or random response pattern. Participants who were deemed to have no response pattern, but whose slopes were deemed to reflect this (i.e., were less than around 0.2), were included. The models predicting BOUNDARY LOCATION included all participants who exhibited either the expected or opposite response pattern (according to the categorization shown in **Figure 5.16**), while the models predicting RESPONSE PATTERN as a categorical outcome variable included all participants, sorted into those who had the expected pattern vs. those who did not (i.e., including those who had either the opposite pattern or no pattern).

Linear models (with the *lmer()* function from the *lme4()* package) were run in R for all outcome variables except for RESPONSE PATTERN, which was coded as a binary categorical response variable and thus required logistic regression models implemented with the *glmer()* function from the *lme4()* package. P-values were computed using the *lmerTest* package. Like the linear models presented in Chapter 4, all models included random intercepts for participants, which were intended to account for individual participants' different baseline performance on the various outcome metrics. Continuous predictors were scaled using the *scale()* function, and collinearity of predictors was tested using the *vif()* function from the *car* package.

Models presented in Section 5.6.4 that included only the four repeated stimuli words generally include the WORD predictor rather than the VOWEL PAIR predictor. These models generally did not test interactions between VOWEL PAIR and demographic factors, production metrics, or AX metrics, as interactions that do not include the full data set (all six stimuli words) are not necessarily interesting or informative. Exceptions were made in those cases where the separate models (for each vowel pair) indicated that some trends may be present among only one of the vowel pairs. In these cases, interactions with VOWEL PAIR were tested in the pooled model in order to verify that the pooled and separate models were indicating the same trends (which justifies not presenting both the pooled and separate models in the results section below).

Models that included all six stimuli words included the VOWEL PAIR predictor rather than the WORD predictor, and tested interactions between VOWEL PAIR and demographic factors,

production metrics, and AX metrics, with the goal of testing whether the relationship of these factors to participants' responses was different across the two vowel pairs.

As with the linear models discussed in Chapter 4, the WORD or VOWEL PAIR predictors were included in all models, as the core goal of the models was to explore patterns observed across the different words and vowel pairs. Other predictors were removed from models when they were clearly not significant, but cases where predictors approached significance will be presented and discussed below, especially because whether or not a particular predictor was significant often depended on the exact analysis that was chosen (i.e., pooled vs. separate model, model with all 6 words vs. the 4 repeated words). AIC and BIC were once again consulted as a general guide in comparing models, but they were not used as a definitive means of choosing which models should be presented and discussed.

5.6.4 Results of linear and logistic regression models

This section presents the results of the linear and logistic regression models discussed in the previous section. As the models discussed in this section tested metrics of performance in the AX task as predictors, Participant 36 (who was excluded from the AX analyses due to very poor performance on the control condition of that task) was excluded from all models. The models presented below thus do not include this participant; however, since the AX metrics were in most cases found to be irrelevant for predicting performance on the dialect categorization task, most of the models presented below were also re-run with the inclusion of Participant 36. In most cases, the inclusion of this participant did not have a significant impact on the results. Significance of some predictors was affected in some models predicting ABSOLUTE VALUE OF SLOPE and some models predicting RESPONSE PATTERN, and the discrepancies between the two sets of models are described in footnotes below. As discussed below, in all cases where the significance of predictors was affected, those predictors had p-values that were already close to 0.05 in the original models. Participant 36 is included in the data visualizations presented below.

Though the WORD or VOWEL PAIR predictor is included in all the models presented below, the significance of such predictors is generally discussed only in the case of the models that included all six stimuli words (rather than only including the four words used as stimuli in both perception tasks). The inclusion of such predictors in the models with the four repeated stimuli words was important for accounting for sources of variation in the given subset of the data. However, drawing conclusions based on comparison of the two vowel pairs in this subset of the

data could be misleading for the purposes of evaluating overall performance on the dialect categorization task, as one third of the data is not included in these models.

5.6.4.1 Predicting SLOPE

In the pooled model that included only the four words used as stimuli in both perception tasks (see **Table 5.19**), BIRTH YEAR was found to be a significant predictor of SLOPE, with younger people having more positive slopes (i.e., slopes that were less consistent with the expected response pattern).

Table 5.19: Results of mixed-effects linear regression model predicting dialect categorization SLOPE for the four stimuli words included in both perception tasks. The reference level of WORD was *mas*. The model included random intercepts for participants. The structure of the model was as follows: **SLOPE ~ WORD + BIRTH YEAR + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places.

Predictor	**p-value**
WORD (*pat*)	*0.012*
WORD (*pos*)	0.729
WORD (*tsots*)	0.571
BIRTH YEAR	*0.015*

In pooled models that included all six stimuli words, a similar trend in relation to BIRTH YEAR was found, though it did not reach significance (p=0.125), while a significant difference between the slopes of the two vowel pairs was found (see **Table 5.20**), with the [a]/[ɒ]$_{A}$ vowel pair having more negative (i.e. more expected) slopes. **Figure 5.17** shows the distribution of slopes by word (the left three words represent [a]/[ɒ]$_{A}$, while the right three words represent [ɔ]/[ɒ]$_{O}$).

Table 5.20: Results of mixed-effects linear regression model predicting dialect categorization SLOPE for all six stimuli words. The reference level of VOWEL PAIR was [a]/[ɒ]$_A$. The model included random intercepts for participants. The structure of the model was as follows: **SLOPE ~ VOWEL PAIR + BIRTH YEAR + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places.

Predictor	**p-value**
VOWEL PAIR	*0.045*
BIRTH YEAR	0.125

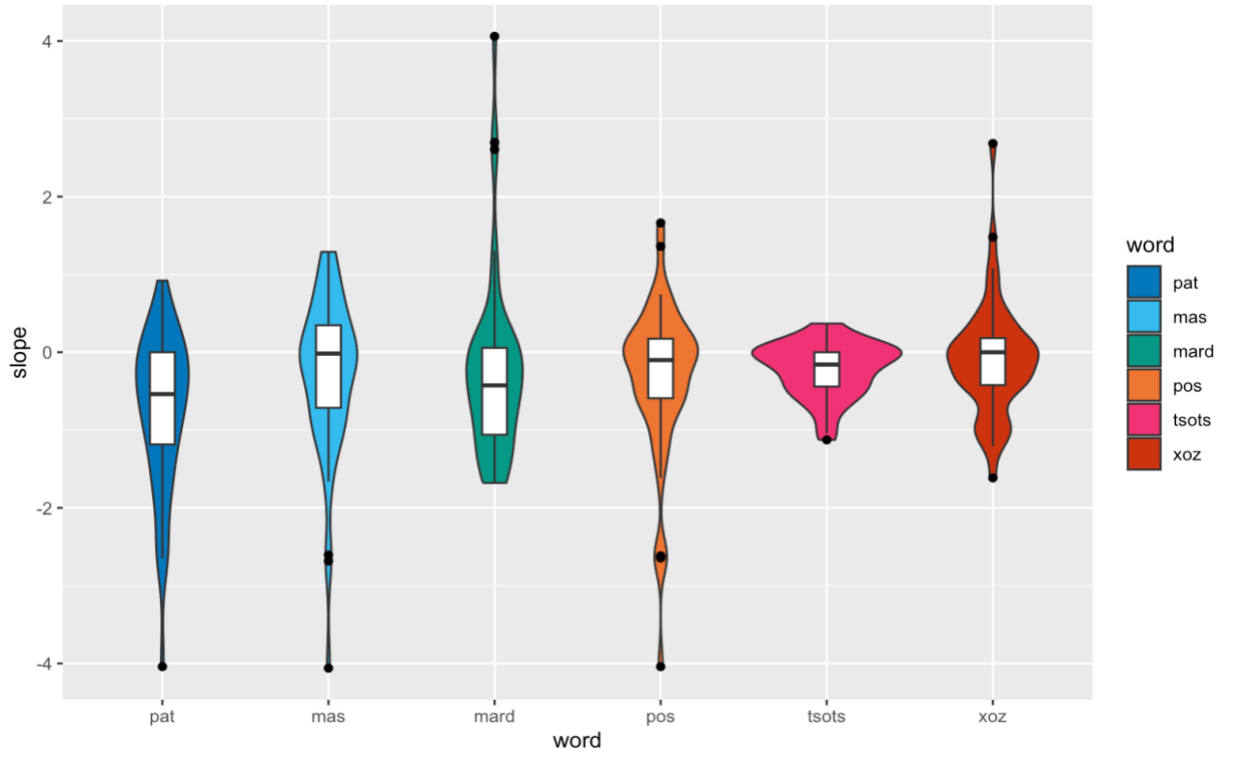


Figure 5.17: Violin plots of participants' individual values of SLOPE by WORD with overlaid boxplots. These plots include those participants who were included in the models predicting slope (see Section 5.6.3). More negative slopes correspond to responding in a more expected manner. The box represents the range between the first and third quartiles (interquartile range), while the horizontal line represents the median. The whiskers represent data points that lie up to 1.5 times the interquartile range from the edge of the box, while black dots represent outliers. The left three words represent [a]/[ɒ]$_A$, while the right three words represent [ɔ]/[ɒ]$_O$.

Significant interactions were also found in pooled models that included all six stimuli words – specifically, both GENDER and BIRTH YEAR interacted significantly with VOWEL PAIR (see **Table 5.21**). These interactions are consistent with the findings of the models that were run separately for each vowel pair (both with only repeated stimuli words and with all six stimuli words), in which a sub-significant gender effect was found for [a]/[ɒ]$_A$ (with men having more positive slopes) and an effect of BIRTH YEAR was found for [ɔ]/[ɒ]$_O$ (with younger people having

more positive slopes). The main effect of BIRTH YEAR in models without interactions is thus likely being driven more so by the [ɔ]/[ɒ]$_O$ vowel pair.

Table 5.21: Results of mixed-effects linear regression model predicting dialect categorization SLOPE for all six stimuli words. The model included random intercepts for participants, as well as interactions of VOWEL PAIR with GENDER and BIRTH YEAR. The structure of the model was as follows: **SLOPE ~ VOWEL PAIR + BIRTH YEAR + GENDER + VOWEL PAIR:BIRTH YEAR + VOWEL PAIR:GENDER + (1 | participant)**. The reference level of VOWEL PAIR was [a]/[ɒ]$_A$. Categorical predictors were dummy coded, such that the coefficients for predictors involved in interactions do not represent main effects. The main effect of VOWEL PAIR was tested using *emmeans()*, and a significant difference was found between the vowel pairs (p=0.041). P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places.

Predictor	**p-value**
VOWEL PAIR	*0.002*
BIRTH YEAR	0.564
GENDER (men vs. women)	*0.034*
VOWEL PAIR – BIRTH YEAR interaction	*0.022*
VOWEL PAIR – GENDER interaction	*0.022*

5.6.4.2 Predicting ABSOLUTE VALUE OF SLOPE

In pooled models that included only the four words used as stimuli in both perception tasks (see **Table 5.22**), GENDER was found to be a significant predictor of ABSOLUTE VALUE OF SLOPE, with men having less steep slopes than women.

Table 5.22: Results of mixed-effects linear regression model predicting ABSOLUTE VALUE OF SLOPE for the four stimuli words included in both perception tasks. The reference level of WORD was *mas.* The model included random intercepts for participants. The structure of the model was as follows: **ABSOLUTE VALUE OF SLOPE ~ WORD + GENDER + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places.

Predictor	**p-value**
WORD (*pat*)	0.221
WORD (*pos*)	0.869
WORD (*tsots*)	*0.002*
GENDER (men vs. women)	*0.007*

Interactions between VOWEL PAIR, GENDER and EDUCATION LEVEL were tested in the pooled models with only repeated stimuli words because of effects that were found in the analogous models that were run separately for each vowel pair. Namely, in the separate model with only repeated [ɔ]/[ɒ]$_O$ stimuli, no demographic predictors were significant, while in the separate model for [a]/[ɒ]$_A$, effects of GENDER and education LEVEL were found (with men and people without higher education having less steep slopes). As the model results presented in **Table 5.23** show, the interaction with EDUCATION LEVEL was nearly significant,[30] while the interaction with GENDER was much farther from significance (p=0.298), suggesting that both vowel pairs nevertheless contribute to the main effect of GENDER found in the model without interactions.

Table 5.23: Results of mixed-effects linear regression model predicting ABSOLUTE VALUE OF SLOPE for the four stimuli words included in both perception tasks. The model included random intercepts for participants, as well as interactions between VOWEL PAIR, GENDER and EDUCATION LEVEL. The structure of the model was as follows: **ABSOLUTE VALUE OF SLOPE ~ VOWEL PAIR + GENDER + EDUCATION LEVEL + VOWEL PAIR:EDUCATION LEVEL + VOWEL PAIR:GENDER + (1 | participant)**. The reference level of VOWEL PAIR was [a]/[ɒ]$_A$. Categorical predictors were dummy coded, such that the coefficients for predictors involved in interactions do not represent main effects. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places.

Predictor	**p-value**
VOWEL PAIR	*<0.001*
GENDER (men vs. women)	*0.015*
EDUCATION LEVEL (non-higher vs. higher)	*0.031*
VOWEL PAIR – EDUCATION LEVEL interaction	0.063
VOWEL PAIR – GENDER interaction	0.298

In pooled models that included all six stimuli words, significant main effects of GENDER and BIRTH YEAR were found (see **Table 5.24**), such that men and younger people had less steep slopes.[31] While similar but sub-significant gender effects were also found in the analogous models run separately for both vowel pairs, the separate model for [a]/[ɒ]$_A$ also had a significant effect of BIRTH YEAR (younger people having less steep slopes) and a sub-significant effect of EDUCATION LEVEL (people without higher education having less steep slopes). This separate model for [a]/[ɒ]$_A$ also had a sub-significant effect of *d'*, in which participants with higher *d'*

[30] In the identical model that included participant 36, this interaction reached significance (p=0.044).
[31] In the identical model that included participant 36, the gender effect did not quite reach significance (p=0.063).

values had steeper slopes (see Chapter 6 for further discussion/visualizations of demographic effects and of comparison of performance across the two tasks). Given these differences between the separate models, interactions of VOWEL PAIR with BIRTH YEAR, EDUCATION LEVEL, and *d'* were tested in the pooled models. As **Table 5.25** shows, only the interaction with EDUCATION LEVEL reached significance. The directionality of the interaction with BIRTH YEAR is consistent with the main effect that was found for this variable and with the results of the models run separately for each vowel pair – in other words, younger people had less steep slopes for both vowel pairs, but this effect was more dramatic for [a]/[ɒ]$_A$. A significant difference between the vowel pairs was also found in both of the presented models, such that [a]/[ɒ]$_A$ had steeper slopes than [ɔ]/[ɒ]$_O$. **Figure 5.18** shows the distribution of ABSOLUTE VALUE OF SLOPE by WORD (the left three words represent [a]/[ɒ]$_A$, while the right three words represent [ɔ]/[ɒ]$_O$).

Table 5.24: Results of mixed-effects linear regression model predicting ABSOLUTE VALUE OF SLOPE for all six stimuli words. The reference level of VOWEL PAIR was [a]/[ɒ]$_A$. The model included random intercepts for participants. The structure of the model was as follows: **ABSOLUTE VALUE OF SLOPE ~ VOWEL PAIR + GENDER + BIRTH YEAR + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places.

Predictor	**p-value**
VOWEL PAIR	*<0.001*
GENDER (men vs. women)	*0.038*
BIRTH YEAR	*0.033*

Table 5.25: Results of mixed-effects linear regression model predicting ABSOLUTE VALUE OF SLOPE for all six stimuli words. The model included random intercepts for participants, as well as interactions of VOWEL PAIR with BIRTH YEAR, EDUCATION LEVEL, and *d'*. The structure of the model was as follows: **ABSOLUTE VALUE OF SLOPE ~ VOWEL PAIR + d' + GENDER + BIRTH YEAR + EDUCATION LEVEL + VOWEL PAIR:d' + VOWEL PAIR:BIRTH YEAR + VOWEL_PAIR:EDUCATION LEVEL + (1 | participant)**. The reference level of VOWEL PAIR was [a]/[ɒ]$_A$. Categorical predictors were dummy coded, such that the coefficients for predictors involved in interactions do not represent main effects. The main effect of VOWEL PAIR was tested using *emmeans()*, and a significant difference was found between the vowel pairs ($p<0.001$). P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places.

Predictor	**p-value**
VOWEL PAIR	*<0.001*
d'	0.119
GENDER (men vs. women)	0.069

BIRTH YEAR	*0.018*
EDUCATION LEVEL (non-higher vs. higher)	0.066
VOWEL PAIR – *d'* interaction	0.226
VOWEL PAIR – BIRTH YEAR interaction	0.128
VOWEL PAIR – EDUCATION LEVEL interaction	*0.049*

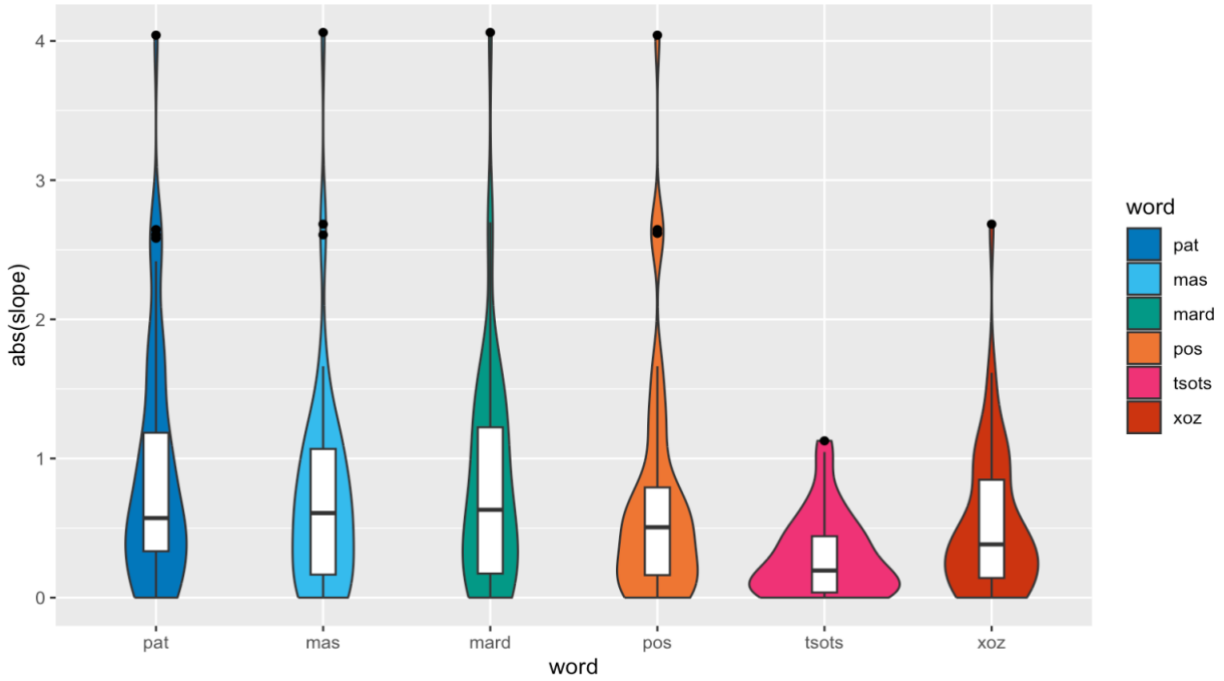


Figure 5.18: Violin plots of ABSOLUTE VALUE OF SLOPE by WORD with overlaid boxplots. These plots include those participants who were included in the models predicting slope (see Section 5.6.3). The box represents the range between the first and third quartiles (interquartile range), while the horizontal line represents the median. The whiskers represent data points that lie up to 1.5 times the interquartile range from the edge of the box, while black dots represent outliers. The left three words represent [a]/[ɒ]$_A$, while the right three words represent [ɔ]/[ɒ]$_O$.

5.6.4.3 Predicting SLOPE only among participants with the expected response pattern

In pooled models that included only the four words used as stimuli in both perception tasks (see **Table 5.26**), both GENDER and PREVIOUS were found to be significant predictors, such that men had more positive slopes (i.e., less steep, as all participants included in the given models had negative slopes), and previous participants had more negative (i.e., steeper) slopes.

Table 5.26: Results of mixed-effects linear regression model predicting SLOPE among participants with the expected response pattern for the four stimuli words included in both perception tasks. The reference level of WORD was *mas*. The model included random intercepts for participants. The structure of the model was as follows: **SLOPE ~ WORD + GENDER + PREVIOUS + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places.

Predictor	p-value
WORD (*pat*)	0.369
WORD (*pos*)	0.854
WORD (*tsots*)	0.054
GENDER (men vs. women)	*0.024*
PREVIOUS	*0.007*

The pooled models that included all six stimuli words showed the same demographic trends, and a significant difference was also found between the two vowel pairs, with [a]/[ɒ]$_A$ having more negative (i.e., steeper) slopes (see **Table 5.27**). **Figure 5.19** visualizes the SLOPE values among participants with the expected pattern for each word.

Table 5.27: Results of mixed-effects linear regression model predicting SLOPE among participants with the expected response pattern for all six stimuli words. The reference level of VOWEL PAIR was [a]/[ɒ]$_A$. The model included random intercepts for participants. The structure of the model was as follows: **SLOPE ~ VOWEL PAIR + GENDER + PREVIOUS + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places.

Predictor	p-value
VOWEL PAIR	*0.012*
GENDER (men vs. women)	*0.025*
PREVIOUS	*0.020*

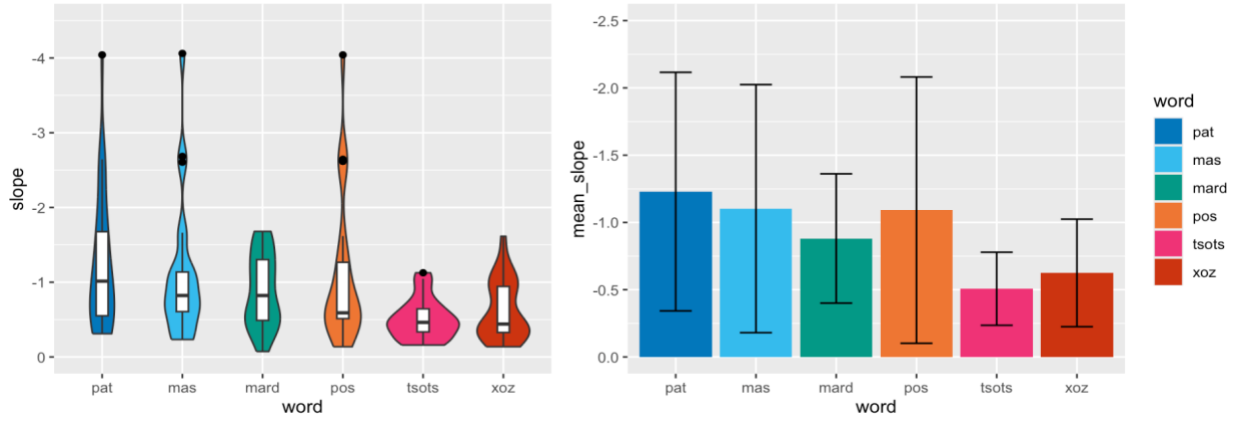


Figure 5.19: The left panel presents violin plots showing the distribution of SLOPE by WORD among only those participants who had the expected response pattern. The box represents the range between the first and third quartiles (interquartile range), while the horizontal line represents the median. The whiskers represent data points that lie up to 1.5 times the interquartile range from the edge of the box, while black dots represent outliers. The right

panel shows mean SLOPE values by WORD among only those participants who had the expected response pattern. Error bars represent ± one standard deviation. The y-axis is inverted, such that higher (more negative) values correspond with steeper slopes. The left three words in each panel represent [a]/[ɒ]$_A$, while the right three words in each panel represent [ɔ]/[ɒ]$_O$.

5.6.4.4 Predicting RESPONSE PATTERN

In pooled models that included only the four words used as stimuli in both perception tasks (see **Table 5.28**), BIRTH YEAR was found to be a significant predictor, with younger people having less of the expected pattern. In the analogous models that were separated by vowel pair, a significant effect of BIRTH YEAR was found for [ɔ]/[ɒ]$_O$ but not for [a]/[ɒ]$_A$. The interaction between VOWEL PAIR and BIRTH YEAR was thus tested in a pooled model (**Table 5.29**), in which the interaction was found to be significant, suggesting that the main effect in the simpler model is driven more so by [ɔ]/[ɒ]$_O$.

Table 5.28: Results of mixed-effects logistic regression model predicting RESPONSE PATTERN for the four stimuli words included in both perception tasks. The reference level of WORD was *mas*. The model included random intercepts for participants. The structure of the model was as follows: **RESPONSE PATTERN ~ WORD + BIRTH YEAR + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places.

Predictor	**p-value**
WORD (*pat*)	0.064
WORD (*pos*)	1.000
WORD (*tsots*)	0.678
BIRTH YEAR	*0.032*

Table 5.29: Results of mixed-effects logistic regression model predicting RESPONSE PATTERN for the four stimuli words included in both perception tasks. The model included random intercepts for participants, as well as the interaction of VOWEL PAIR and BIRTH YEAR. The structure of the model was as follows: **RESPONSE PATTERN ~ VOWEL PAIR * BIRTH YEAR + (1 | participant)**. The reference level of VOWEL PAIR was [a]/[ɒ]$_A$. Categorical predictors were dummy coded, such that the coefficients for predictors involved in interactions do not represent main effects. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places.

Predictor	**p-value**
VOWEL PAIR	0.103
BIRTH YEAR	0.849
VOWEL PAIR – BIRTH YEAR interaction	*0.045*

In the pooled models that included all six stimuli words, a sub-significant effect of BIRTH YEAR was also found (p=0.079), with younger people having less of the expected pattern (see **Table 5.30**). A significant difference between the vowel pairs was also found in the model presented in **Table 5.30**, with [a]/[ɒ]$_A$ having more of the expected pattern.[32]

Table 5.30: Results of mixed-effects logistic regression model predicting RESPONSE PATTERN for all six stimuli words. The reference level of VOWEL PAIR was [a]/[ɒ]$_A$. The model included random intercepts for participants. The structure of the model was as follows: **RESPONSE PATTERN ~ VOWEL PAIR + BIRTH YEAR + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places.

Predictor	**p-value**
VOWEL PAIR	*0.040*
BIRTH YEAR	0.079

Significant interactions between VOWEL PAIR and GENDER and VOWEL PAIR and BIRTH YEAR were found (see **Table 5.31** and **Table 5.33**). These interactions indicated that the age effect shown in **Table 5.30** was being driven primarily by [ɔ]/[ɒ]$_O$, while the performance of the two gender groups (in terms of which group had more of the expected pattern) was different across the two vowel pairs (with men having less of the expected pattern for [a]/[ɒ]$_A$ and more of it for [ɔ]/[ɒ]$_O$).

Sub-significant interactions with two metrics from the AX task were also found – ACCURACY and *d'* (see **Tables 5.31** and **5.33**). However, as the slopes for each vowel pair in relation to these metrics presented in **Tables 5.32** and **5.34** show, the relationships between AX performance and RESPONSE PATTERN for each vowel pair were opposite for the two metrics. For example, in the case of ACCURACY, participants with higher accuracy showed less of the expected response pattern for [a]/[ɒ]$_A$, but in the case of *d'*, participants with higher *d'* values (which indicate more accurate discrimination) showed more of the expected response pattern for [a]/[ɒ]$_A$. Based on the confidence intervals presented in **Table 5.34**, the effect in relation to *d'* appears more likely to be meaningful (its confidence intervals do not contain 0, suggesting that its slope is significantly different from 0). Nevertheless, the disagreement of the trends in relation to the two metrics is puzzling. While it is thus difficult to interpret these results, they demonstrate that methodological choices (such as whether to evaluate AX performance with raw

[32] In the identical model that included participant 36, VOWEL PAIR was not quite significant (p=0.057).

proportions or with *d'*) can affect the results of analyses (see Chapter 6 for visualizations of these effects and further discussion of performance across the two perception tasks).

A significant difference between the vowel pairs was also found in both models presented in **Table 5.31** and **Table 5.33** (with [a]/[ɒ]$_A$ having more of the expected pattern). **Figure 5.20** below shows the number of participants with the expected pattern for each word.

Table 5.31: Results of mixed-effects logistic regression model predicting RESPONSE PATTERN for all six stimuli words. The model included random intercepts for participants, as well as interactions of VOWEL PAIR with BIRTH YEAR, GENDER, and AX ACCURACY. The structure of the model was as follows: **RESPONSE PATTERN ~ VOWEL PAIR + GENDER + BIRTH YEAR + ACCURACY + VOWEL PAIR:GENDER + VOWEL PAIR:BIRTH YEAR + VOWEL PAIR:ACCURACY + (1 | participant)**. The reference level of VOWEL PAIR was [a]/[ɒ]$_A$. Categorical predictors were dummy coded, such that the coefficients for predictors involved in interactions do not represent main effects. The main effect of VOWEL PAIR was tested using *emmeans()*, and a significant difference was found between the vowel pairs (p=0.028). P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places.

Predictor	**p-value**
VOWEL PAIR	*0.002*
GENDER (men vs. women)	0.061
BIRTH YEAR	0.258
ACCURACY	0.508
VOWEL PAIR – GENDER interaction	*0.016*
VOWEL PAIR – BIRTH YEAR interaction	*<0.001*
VOWEL PAIR – ACCURACY interaction	0.121

Table 5.32: Slopes of AX ACCURACY across vowel pairs from the model presented in **Table 5.31**, computed using *emtrends()*. 0.95 confidence intervals calculated with *emtrends()* are provided.

Vowel pair	**Slope**	**Lower CI**	**Upper CI**
[a]/[ɒ]$_A$	0.136	-0.267	0.540
[ɔ]/[ɒ]$_O$	-0.287	-0.698	0.123

Table 5.33: Results of mixed-effects logistic regression model predicting RESPONSE PATTERN for all six stimuli words. The model included random intercepts for participants, as well as interactions of VOWEL PAIR with BIRTH YEAR, GENDER, and *d'*. The structure of the model was as follows: **RESPONSE PATTERN ~ VOWEL PAIR + GENDER + BIRTH YEAR + d' + VOWEL PAIR:GENDER + VOWEL PAIR:BIRTH YEAR + VOWEL PAIR:d' + (1 | participant)**. The reference level of VOWEL PAIR was [a]/[ɒ]$_A$. Categorical predictors were dummy coded, such that the coefficients for predictors involved in interactions do not represent main effects. The main effect of VOWEL PAIR was tested using

emmeans(), and a significant difference was found between the vowel pairs (p=0.049). P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places.

Predictor	**p-value**
VOWEL PAIR	*0.002*
GENDER (men vs. women)	*0.045*
BIRTH YEAR	0.067
d'	*0.038*
VOWEL PAIR – GENDER interaction	*0.015*
VOWEL PAIR – BIRTH YEAR interaction	*<0.001*
VOWEL PAIR – *d'* interaction	0.080

Table 5.34: Slopes of *d'* across vowel pairs from the model presented in **Table 5.33**, computed using *emtrends()*. 0.95 confidence intervals calculated with *emtrends()* are provided.

Vowel pair	**Slope**	**Lower CI**	**Upper CI**
[a]/[ɒ]$_A$	-0.5008	-0.973	-0.028
[ɔ]/[ɒ]$_O$	0.0227	-0.359	0.404

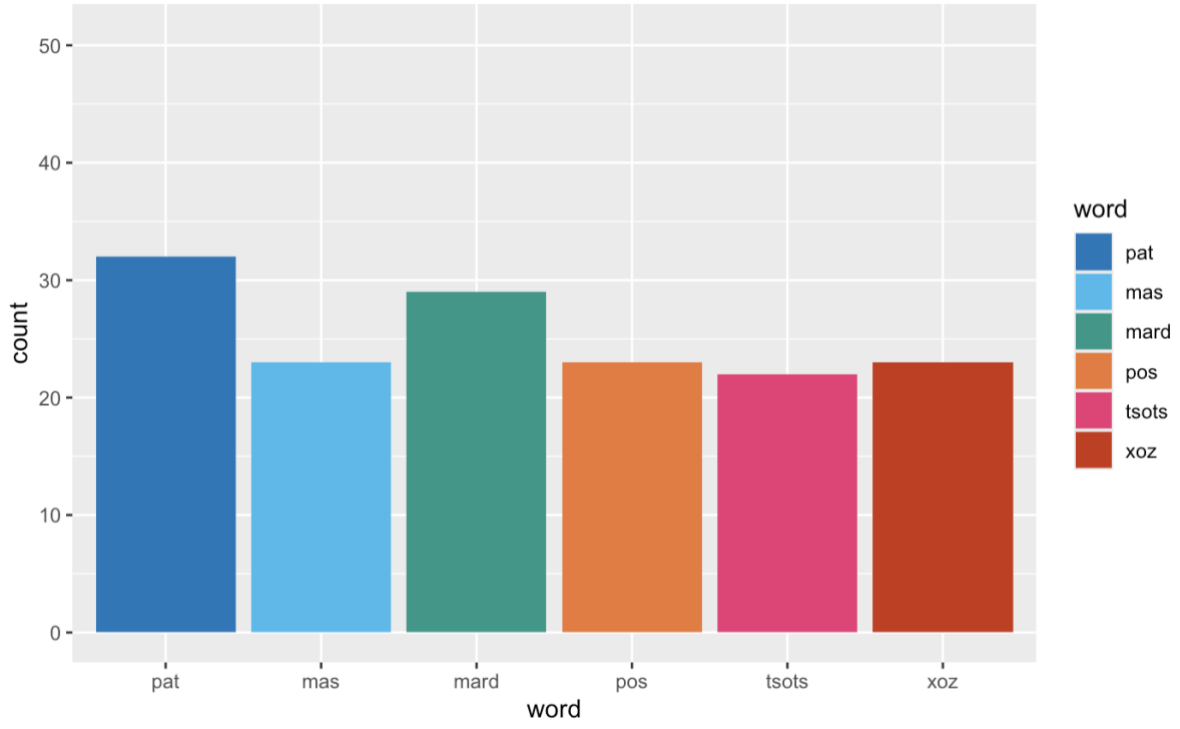


Figure 5.20: Number of participants with the expected response pattern by WORD. The word *pat* had the highest number of expected patterns (N=32). The left three words represent [a]/[ɒ]$_A$, while the right three words represent [ɔ]/[ɒ]$_O$.

5.6.4.5 Predicting BOUNDARY LOCATION

In pooled models that included only the four words used as stimuli in both perception tasks, main effects were not found for any demographic, production, or AX predictors. Significant differences were found only in relation to WORD, such that *pos* and *tsots* had boundary locations that were significantly further toward the right in comparison with the reference level *mas*, while the difference between *pat* and *mas* was not significant (see **Table 5.35**). Several interactions with VOWEL PAIR were tested because of effects that were found in the analogous models separated by vowel pair. In the separate model for [ɔ]/[ɒ]$_O$, no predictors were significant, while in the separate model for [a]/[ɒ]$_A$, BIRTH YEAR was significant (with younger people having boundaries more toward the left), and Pillai scores and ED-median (from non-Bark-converted data) were sub-significant (with people with higher scores / distance [i.e., more dialect-like vowels] having boundaries more toward the left). The model presented in **Table 5.36** below includes only the interactions with BIRTH YEAR and ED-median, as ED-median and Pillai scores are somewhat collinear and also had effects of similar magnitude and direction relative to the response variable (though the interaction with ED-median was slightly closer to significance). As the table shows, the interaction with BIRTH YEAR was significant, while the interaction with ED-median was not.

Table 5.35: Results of mixed-effects linear regression model predicting BOUNDARY LOCATION for the four stimuli words included in both perception tasks. The reference level of WORD was *mas.* The model included random intercepts for participants. The structure of the model was as follows: **BOUNDARY LOCATION ~ WORD + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places.

Predictor	**p-value**
WORD (*pat*)	0.558
WORD (*pos*)	*0.018*
WORD (*tsots*)	*0.011*

Table 5.36: Results of mixed-effects linear regression model predicting BOUNDARY LOCATION for the four stimuli words included in both perception tasks. The model included random intercepts for participants, as well as the interactions of VOWEL PAIR with BIRTH YEAR and ED-median (from non-Bark-converted data). The structure of the

model was as follows: **BOUNDARY LOCATION ~ VOWEL PAIR + ED-median + BIRTH YEAR + VOWEL PAIR:BIRTH YEAR + VOWEL PAIR:ED-median + (1 | participant)**. The reference level of VOWEL PAIR was [a]/[ɒ]$_A$. Categorical predictors were dummy coded, such that the coefficients for predictors involved in interactions do not represent main effects. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places.

Predictor	**p-value**
VOWEL PAIR	0.058
BIRTH YEAR	*0.047*
ED-median	0.309
VOWEL PAIR – BIRTH YEAR interaction	*0.007*
VOWEL PAIR – ED-median interaction	0.320

In the pooled models that included all six stimuli words, a sub-significant effect (p=0.059) of EDUCATION LEVEL was found (see **Table 5.37**), such that people without higher education had boundaries more toward the right. The difference between the vowel pairs was also significant in this model, with [ɔ]/[ɒ]$_O$ words having boundaries mores toward the right and [a]/[ɒ]$_A$ words having boundaries more toward the left. While the different vowel pairs having different boundary locations is not necessarily unexpected or inherently interesting (given that the two continua represent different areas of the vowel space), the differing boundary locations have potential implications when considered together with the results of the AX task, as further discussed in Section 5.7.2.1.

Because of effects found in the analogous separate models by vowel pair, several interactions were tested. In the separate model for [ɔ]/[ɒ]$_O$, no predictors were significant, while in the separate model for [a]/[ɒ]$_A$, EDUCATION LEVEL and BIRTH YEAR were significant (with people with higher education and younger people having boundaries more toward the left), and Pillai scores and ED-median (from non-Bark-converted data) were sub-significant (with people with higher scores / distance (i.e., more dialect-like vowels) having boundaries more toward the left). The model presented in **Table 5.38** below includes only the interactions with BIRTH YEAR, EDUCATION LEVEL, and ED-median (rather than Pillai scores, for the same reasons listed above in relation to the models with only repeated stimuli words). The interaction with BIRTH YEAR was significant, indicating that the relationship between BIRTH YEAR and BOUNDARY LOCATION is different for the two vowel pairs. The interaction with ED-median was somewhat close to

significant (p=0.152), while the interaction with EDUCATION LEVEL was not, suggesting that both vowel pairs contribute to the main effect of EDUCATION LEVEL found in the simpler model.

In the model with interactions, the difference between the vowel pairs is still close to significant (p=0.051), and the directionality of the difference is the same as in the simpler model. In both models, boundary locations between steps 3-4 (3.42 and 3.8) are predicted for [a]/[ɒ]$_A$, while for [ɔ]/[ɒ]$_O$, the simple model predicts a boundary location of 4.9, and the model with interactions predicts a boundary location of 5.9. **Figure 5.21** visualizes the distribution of individual boundary locations by WORD.

Table 5.37: Results of mixed-effects linear regression model predicting BOUNDARY LOCATION for all six stimuli words. The reference level of VOWEL PAIR was [a]/[ɒ]$_A$. The model included random intercepts for participants. The structure of the model was as follows: **BOUNDARY LOCATION ~ VOWEL PAIR + EDUCATION LEVEL + (1 | participant)**. P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places.

Predictor	**p-value**
VOWEL PAIR	*<0.001*
EDUCATION LEVEL (non-higher vs. higher)	0.059

Table 5.38: Results of mixed-effects linear regression model predicting BOUNDARY LOCATION for all six stimuli words. The model included random intercepts for participants, as well as interactions of VOWEL PAIR with BIRTH YEAR, EDUCATION LEVEL, and ED-median (calculated from non-Bark-converted data). The structure of the model was as follows: **BOUNDARY LOCATION ~ VOWEL PAIR + BIRTH YEAR + EDUCATION LEVEL + ED-median + VOWEL PAIR:BIRTH YEAR + VOWEL PAIR:EDUCATION LEVEL + VOWEL PAIR:ED-median + (1 | participant)**. The reference level of VOWEL PAIR was [a]/[ɒ]$_A$. Categorical predictors were dummy coded, such that the coefficients for predictors involved in interactions do not represent main effects. The main effect of VOWEL PAIR was tested using *emmeans()*, and a nearly significant difference was found between the vowel pairs (p=0.051). P-values of less than 0.05 are italicized. All values are rounded to 3 decimal places.

Predictor	**p-value**
VOWEL PAIR	*0.048*
BIRTH YEAR	*0.022*
EDUCATION LEVEL (non-higher vs. higher)	0.059
ED-median	0.161
VOWEL PAIR – BIRTH YEAR interaction	*0.006*
VOWEL PAIR – EDUCATION LEVEL interaction	0.727

VOWEL PAIR – ED-median interaction	0.152

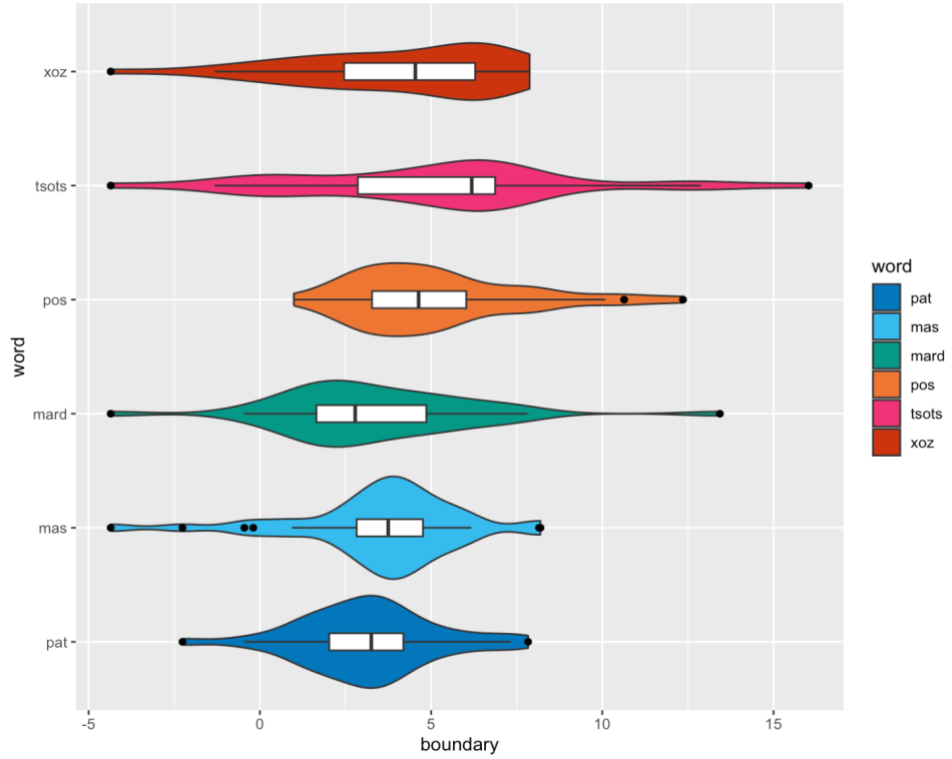


Figure 5.21: Violin plots of BOUNDARY LOCATION by WORD with overlaid boxplots. These plots include those participants who were deemed to have either the expected or opposite response pattern. Values toward the left of the plot correspond with boundary locations farther toward the dialect-like end of the continuum, while values toward the right of the plot correspond with boundary locations farther toward the SEA-like end of the continuum. As shown in the plot and noted in Section 5.6.1, some participants' boundary locations fell outside of the values of the continuum steps (i.e., were less than 1 or greater than 7). The box represents the range between the first and third quartiles (interquartile range), while the horizontal line represents the median. The whiskers represent data points that lie up to 1.5 times the interquartile range from the edge of the box, while black dots represent outliers. The bottom three words represent [a]/[ɒ]$_A$, while the top three words represent [ɔ]/[ɒ]$_O$.

5.6.4.6 Summary of results of individual-level analyses

Tables 5.39-5.43 below summarize the results of the models presented in Sections 5.6.4.1-5.6.4.5, including information about demographic, production, and AX predictors. Similarly to the sections above, the tables below only note the significance of the VOWEL PAIR predictor in the models that included all 6 stimuli words, as these models reflect the entire data set of the task.

As the tables show, VOWEL PAIR was almost always significant in the models that included all 6 stimuli words, with the only exception being the model with interactions predicting BOUNDARY LOCATION (where it was close to significant; p=0.051). In the case of

SLOPE, ABSOLUTE VALUE OF SLOPE, SLOPE among participants with the expected pattern, and RESPONSE PATTERN, responses for [a]/[ɒ]$_A$ conformed more closely with expectations about performance on the task than responses for [ɔ]/[ɒ]$_O$. The [a]/[ɒ]$_A$ vowel pair had slopes that were more negative (i.e., more expected) and steeper both in general and among participants with the expected response pattern. This vowel pair also had more of the expected response pattern when participants were classified categorically. Thus, the results of the analyses presented in Sections 5.6.4.1-5.6.4.5 align with those presented in Section 5.5 in suggesting that participants generally responded in a more expected manner for the socially salient [a]/[ɒ]$_A$ pair. The models also found significant or nearly significant differences in BOUNDARY LOCATION for the two vowel pairs, such that boundaries for [a]/[ɒ]$_A$ were estimated to be near the middle of the continuum, while boundaries for [ɔ]/[ɒ]$_O$ fell more toward the right. This aligns with the general expectation that the crossover point from hearing dialect to non-dialect variants could differ for the two vowel pairs. The finding that [a]/[ɒ]$_A$ boundaries were further to the left also aligns with the findings of the AX task, according to which discrimination of this vowel pair was poor toward the right (SEA-like) end of the continuum.

Table 5.39: Summary of key results of models predicting SLOPE.

	4 repeated stimuli words	**All 6 stimuli words**	**All 6 stimuli words with interactions**
significant	BIRTH YEAR (younger people have more positive [less expected] slopes)	VOWEL PAIR ([a]/[ɒ]$_A$ has more negative [more expected] slopes)	VOWEL PAIR ([a]/[ɒ]$_A$ has more negative [more expected] slopes); VOWEL PAIR – BIRTH YEAR interaction (younger people have more positive slopes especially for [ɔ]/[ɒ]$_O$); VOWEL PAIR – GENDER interaction (men have more positive slopes for [a]/[ɒ]$_A$).
sub-significant		BIRTH YEAR (younger people have more positive [less expected] slopes)	

Table 5.40: Summary of key results of models predicting ABSOLUTE VALUE OF SLOPE.

	4 repeated stimuli words	**4 repeated stimuli words with interactions**	**All 6 stimuli words**	**All 6 stimuli words with interactions**
significant	GENDER (men have less steep slopes)		VOWEL PAIR ([a]/[ɒ]$_A$ has steeper slopes); GENDER (men have less steep slopes); BIRTH YEAR	VOWEL PAIR ([a]/[ɒ]$_A$ has steeper slopes); VOWEL PAIR – EDUCATION LEVEL

			(younger people have less steep slopes)	interaction (people without higher education have less steep slopes for [a]/[ɒ]$_A$)
sub-significant		VOWEL PAIR – EDUCATION LEVEL interaction (people without higher education have less steep slopes for [a]/[ɒ]$_A$)		VOWEL PAIR – BIRTH YEAR interaction (younger people had less steep slopes for both vowel pairs, but effect more dramatic for [a]/[ɒ]$_A$)
Otherwise notable				VOWEL PAIR – *d'* interaction not very close to significant (p=0.226), but in separate model for [a]/[ɒ]$_A$, there was sub-significant effect of *d'* (higher *d'* = steeper slopes).

Table 5.41: Summary of key results of models predicting SLOPE only among participants with expected response pattern.

	4 repeated stimuli words	**All 6 stimuli words**
significant	GENDER (men have less steep slopes); PREVIOUS (previous participants have steeper slopes)	VOWEL PAIR ([a]/[ɒ]$_A$ has steeper slopes); GENDER (men have less steep slopes); PREVIOUS (previous participants have steeper slopes)
sub-significant		

Table 5.42: Summary of key results of models predicting RESPONSE PATTERN.

	4 repeated stimuli words	**4 repeated stimuli words with interactions**	**All 6 stimuli words**	**All 6 stimuli words with interactions**
significant	BIRTH YEAR (younger people have less of expected pattern)	VOWEL PAIR – BIRTH YEAR interaction (main effect of birth year in the simpler model is driven more by [ɔ]/[ɒ]$_O$)	VOWEL PAIR ([a]/[ɒ]$_A$ has more of expected pattern)	VOWEL PAIR ([a]/[ɒ]$_A$ has more of expected pattern); VOWEL PAIR – GENDER interaction (men having less of the expected pattern for [a]/[ɒ]$_A$ and more of it for [ɔ]/[ɒ]$_O$); VOWEL PAIR – BIRTH YEAR interaction (main effect of birth year in the simpler model is driven more by [ɔ]/[ɒ]$_O$)
sub-significant			BIRTH YEAR (younger people have less of expected pattern)	VOWEL PAIR – ACCURACY interaction (higher accuracy = less of expected response pattern for [a]/[ɒ]$_A$); VOWEL PAIR – *d'*

				interaction (higher *d'* [higher accuracy] = more of expected response pattern for [a]/[ɒ]$_A$)

Table 5.43: Summary of key results of models predicting BOUNDARY LOCATION.

	4 repeated stimuli words	**4 repeated stimuli words with interactions**	**All 6 stimuli words**	**All 6 stimuli words with interactions**
significant		VOWEL PAIR – BIRTH YEAR interaction (for [a]/[ɒ]$_A$ younger people had boundaries more toward the left)	VOWEL PAIR ([a]/[ɒ]$_A$ has boundaries more toward the left);	VOWEL PAIR – BIRTH YEAR interaction (relationship between age and boundary location is different for two vowel pairs)
sub-significant			EDUCATION LEVEL (people without higher education had boundaries more toward the right)	VOWEL PAIR ([a]/[ɒ]$_A$ has boundaries more toward the left); VOWEL PAIR – ED-median interaction (more dialect like vowels = boundary more toward left for [a]/[ɒ]$_A$)
Otherwise notable		VOWEL PAIR – ED-median interaction not very close to significant (p=0.320), but in separate models for [a]/[ɒ]$_A$, there was sub-significant effect of ED-median and Pillai scores, such that people with more dialect-like vowels had boundaries more toward the left		

5.7 General Discussion

5.7.1 Group-level logistic regression models

5.7.1.1 Comparison of results for the two vowel pairs

According to the results of the large logistic regression models presented in Section 5.5, the expected response pattern (more "Gavaṛ" responses toward the left end of the continuum and more "not Gavaṛ" responses toward the right end of the continuum) was exhibited at the group level for all six stimuli words. Nevertheless, comparison of the visualizations of pooled data presented in **Figures 5.5** and **5.8** raises questions about more subtle differences in performance across the two vowel pairs. In particular, in these pooled visualizations, all three words representing [ɔ]/[ɒ]$_O$ show a strange pattern in which the amount of "Gavaṛ" responses decreases as the steps get closer to the right end of the continuum, only to noticeably increase again at step

7. In contrast, none of the words representing [a]/[ɒ]$_A$ showed this pattern. The pooled raw data for [a]/[ɒ]$_A$ thus seems more consistent with the expected response pattern and with the negative slopes estimated by the models, while the pooled raw data for [ɔ]/[ɒ]$_O$ seems like it could contain more substantial individual variation or noise in response patterns. Examination of the estimated slopes and confidence intervals of the STEP predictor for each word (**Tables 5.12** and **5.16**) lends further support to the observation that responses for [a]/[ɒ]$_A$ may have been more in line with the expected pattern. While both vowel pairs had only one word whose confidence intervals contained zero (suggesting a slope that is not significantly different from zero), the upper confidence interval for the [a]/[ɒ]$_A$ word (*mas*) was just barely above zero (0.041), while [ɔ]/[ɒ]$_O$ had one word (*xoz*) whose upper confidence interval was substantially higher than zero (0.267), and one (*pos*) whose upper confidence interval was negative but nevertheless very close to zero (-0.008). Thus, while the differences between responses to the two vowel pairs may be subtle, the [a]/[ɒ]$_A$ pair seems to have more significantly negative slopes, which are more in line with the expected response pattern of having a clear difference in responses between the two ends of the continuum.

5.7.1.2 Demographic factors and production metrics

Some weak (most of the time sub-significant) trends in relation to demographic factors and production metrics were also found in these models, but they are more difficult to interpret than the effects in relation to WORD and STEP. In particular, opposite trends in relation to production metrics were found for the two vowel pairs. For [a]/[ɒ]$_A$, participants with greater distances between the relevant vowel categories (i.e., more dialect-like pronunciations) chose "Gavaṛ" less, while for [ɔ]/[ɒ]$_O$, those with less overlap (i.e., more dialect-like pronunciations) chose "Gavaṛ" more. Since the outcome variable of these models was simply the response to each step ("Gavaṛ" vs. "not Gavaṛ"), rather than some more global metric of performance on the task, it is not possible to say if these trends have any relation to the categoricity of responses, to having the expected response pattern, etc., which makes them even more difficult to interpret.

5.7.2 Linear and logistic regression models predicting individual metrics

5.7.2.1 Comparison of results for the two vowel pairs

The results of the individual-level analysis (via the visualizations and models presented in Section 5.6) support the tendency that emerged from the pooled models and visualizations,

according to which responses for [a]/[ɒ]$_A$ were more in line with the expected pattern. This vowel pair had significantly or nearly significantly more expected (i.e., more negative) SLOPE values, steeper SLOPE values (both in general and among those with the expected response pattern), and more of the expected response pattern (in models with the categorical RESPONSE PATTERN outcome variable). Steeper slopes (especially among those with the expected response pattern) correspond to more expected performance on the task, because having a steeper slope indicates having perceived the stimuli in a more categorical manner (i.e., consistently responding "Gavaɻ" at one end of continuum and "not Gavaɻ" at the other). Significant or nearly significant differences in BOUNDARY LOCATION were also found, with boundaries for [a]/[ɒ]$_A$ being close to the middle of the continuum on average, with boundaries for [ɔ]/[ɒ]$_O$ being closer to the right end (SEA-like end) of continuum. The fact that boundaries for [a]/[ɒ]$_A$ were on average not near the right end of the continuum seems consistent with the results of the AX task, in which participants had worse discrimination for this vowel pair at the right end of the continuum, where differences are closer to the prototypical /a/ vowel and are more likely to be perceived as coarticulatory. In contrast, discrimination of [ɔ]/[ɒ]$_O$ was slightly or substantially more accurate toward the right end of the continuum (depending on the stimuli word). Thus, for both vowel pairs, there seems to be some connection between performance on the AX task and boundary location in the dialect categorization task, such that boundary locations do not tend to fall in areas of relatively poor discrimination.

5.7.2.2 Demographic factors, production metrics, and AX metrics

Several demographic trends emerged repeatedly across the models predicting various individual-level metrics. Specifically, younger people performed in a less expected manner according to many metrics (SLOPE, ABSOLUTE VALUE OF SLOPE, RESPONSE PATTERN) as did men (according to SLOPE, ABSOLUTE VALUE OF SLOPE, and SLOPE among those with the expected pattern). People without higher education had less steep slopes (ABSOLUTE VALUE OF SLOPE) for [a]/[ɒ]$_A$, though it is difficult to interpret a trend found in relation to this single metric, since it included both expected and opposite slopes (and thus tells us more about categoricity of response pattern in general, rather than about responding in the expected manner). Several demographic factors were also relevant for predicting BOUNDARY LOCATION, though it is also difficult to say anything about task performance (i.e., about responding in an expected or unexpected manner) based on this metric. The tendency for younger people to have overall worse performance on the

task than older people is consistent with the results of the AX task. In the case of the dialect categorization task, this tendency could be related to younger people having had less lifetime exposure to dialect-like speech (either due to ongoing dialect leveling or simply to their age). This tendency is relatively consistent with the findings of a dialect classification task carried out by Dossey et al. (2020), in which peak performance was found among adults around age 40 (see graph presented by Dossey et al. [2020: 341]), while performance dropped off sharply only after age 50 (i.e., among participants who were older than any participants of the present study). Men's overall worse performance on the task than women creates an interesting contrast with the gender patterns found in production, as the production and perception findings together suggest that men have more dialect-like vowels in production, but are worse at identifying such vowels as dialectal in perception. In any case, the variability of the significance of demographic predictors across different models (e.g., models that were pooled vs. separate for the two vowel pairs) aligns with the analogous analysis of the AX task in suggesting that such methodological decisions can significantly affect results, making it crucial to compare the results of different methods and avoid making arbitrary methodological decisions.

Several trends related to AX and production metrics were also found in these models (see Chapter 6 for visualizations and further discussion of some of these trends). In models predicting ABSOLUTE VALUE OF SLOPE for [a]/[ɒ]$_A$, participants with higher values of *d'* had sub-significantly steeper slopes. Thus, there may be some connection between accurate AX discrimination and perceiving stimuli more categorically in the dialect task. Nevertheless, this effect was very weak (sub-significant), and such a tendency in relation only to ABSOLUTE VALUE OF SLOPE tells us only about categoricity of the response in general, rather than about responding to the dialect categorization task in an expected vs. unexpected manner. In models predicting RESPONSE PATTERN, AX ACCURACY and *d'* related to RESPONSE PATTERN for the two vowel pairs in opposite ways, which is a strange finding given that these metrics are both measures of accuracy of AX discrimination. The meaningfulness of these effects thus seems dubious, and the discrepancy between the results in relation to the two different metrics again highlights the fact that different analytical choices (e.g., which metric to use in evaluating AX performance) can lead to different and even opposing results.

For [a]/[ɒ]$_A$, there was also a sub-significant tendency for people with more dialect-like vowels (greater distances, less overlap between the relevant vowels in production) to have

perception boundaries further toward the left (more dialect-like) end of the continuum. Like the other trends found in relation to production metrics, this trend was very weak and is difficult to interpret. However, perhaps it could indicate a perception-production link, according to which people with more distinct vowel categories in production needed a larger difference from the standard variant in order to perceive the dialect variant. This result and interpretation is also somewhat consistent with the result from the group-level models mentioned above, in which participants with more dialect-like pronunciations of [a]/[ɒ]$_A$ chose "Gavaṛ" significantly less overall in categorizing this pair. This result suggests that participants with more extreme dialect-like pronunciations of this pair might also only perceive the most extreme pronunciations of [ɒ]$_A$ as dialectal (leading to lower overall "Gavaṛ" responses), thus suggesting a perception-production link similar to the one emerging from the analysis of BOUNDARY LOCATION. Caution is nevertheless needed in interpreting this result, not only because of the questionable appropriateness of BOUNDARY LOCATION as a way of quantifying performance on the given task (see Section 5.6.1), but also because the models predicting this metric included participants with both expected and opposite response patterns. While response patterns that were opposite of the expected direction were relatively few for this vowel pair (see Figure 5.16), this means that the analysis of BOUNDARY LOCATION tells us about the location of the perceptual boundary itself, and not necessarily about which endpoint each participant categorized as dialectal. Even so, as the additional figures provided in Appendix F show, the tendency for participants with more dialect-like productions to have further left perception boundaries for [a]/[ɒ]$_A$ was driven primarily by participants with the expected response pattern, lending credibility to the interpretation of a link between dialect-like productions and perception of only larger differences as dialectal.

5.7.3 Discussion of overall task performance

Overall, the most important result of this task (based on both the group and individual-level analyses) is the difference between the results for the two vowel pairs, in accordance with which responses for [a]/[ɒ]$_A$ aligned more closely with expectations than responses for [ɔ]/[ɒ]$_O$. This difference aligns with the differing characteristics of the two vowel pairs – namely, that [a]/[ɒ]$_A$ is more socially salient as a dialect feature than [ɔ]/[ɒ]$_O$. Nevertheless, as **Figures 5.15** and **5.16** show, there was a great deal of individual variation in performance on this task, and even for the word that had the best performance (*pat)*, there were still 18 (out of 50) participants who did not show the expected response pattern. This deviated from the expectations based on

piloting of the task, in which 4 out of 5 pilot participants showed the expected pattern for all three [a]/[ɒ]$_A$ words. (The results are overall nevertheless consistent with the results of piloting in terms of performance across the two vowel pairs, as only 1 out of 5 pilot participants had the expected pattern for [ɔ]/[ɒ]$_O$ words. It is also notable that the most common pattern across the pilot and actual participants was the same – namely, having the expected pattern for all [a]/[ɒ]$_A$ words and no [ɔ]/[ɒ]$_O$ words.)

The overall worse-than-expected performance on this task is likely related to difficulties associated with perceiving words in isolation and perceiving words with extreme (e.g., very dialect-like) variants of vowels. Labov (2010: 59-85) describes a set of experiments in which listeners who were local and non-local to several US cities had to try to recognize words containing the most advanced forms of ongoing sound changes. These experiments found that recognition of isolated words was low overall (2010: 62), and that local listeners had only a small advantage over non-local listeners, often failing to recognize forms that were phonetically similar to their own speech (2010: 84). Labov (2010: 85) emphasizes that hearing words in isolation in these experiments could have contributed to the high error rate, as this is not how words are perceived in everyday speech. Importantly, it seems that a lexical identification task (i.e., identifying what lexical item an isolated, very sound-changed pronunciation corresponds to) is simpler than what participants were asked to do in the present study (i.e., access and use metalinguistic knowledge about dialects in order to sort variants with varying vowels, heard in isolation from any speech context). Thus, it is not very surprising that many participants struggled with the dialect categorization task and did not show the expected response pattern, despite often having large differences between the given vowels in their own speech. There was even some evidence that some participants, similarly to participants of the experiments described by Labov, struggled with lexical recognition of dialectal variants in the task (which may have been more of an issue for some of the stimuli words than for others). For example, one participant, after having responded to the stimuli for the word *pat* in the expected manner, appeared confused when he reached the block for *mas* and heard the dialectal pronunciation ([mɒs]) for the first time, exclaiming, "*ˈmɒsˈn intʃʰ a*?" ("what is *mɒs*?").

There are also additional factors whose influence on individual variation in response patterns is difficult to quantify. For example, it is possible that the particular order chosen for the presentation of the test stimuli (see Section 5.3.3) facilitated categorization for some participants

but not others. Additionally, it is possible that the vowel differences presented in the stimuli were simply not extreme enough to activate listeners' associations with dialect speech. As noted in Chapter 1, Katvalyan (2016) writes that the rounding of [ɒ]$_{A}$ is less extreme in Gavaṛ than in surrounding villages (especially Sarukhan village). As the talker's [a] and [ɒ]$_{A}$ pronunciations were very distinct from one another, but the [ɒ]$_{A}$ pronunciation was not strongly diphthongized, it is possible that this pronunciation was not sufficiently aligned with stereotypes about this vowel. The poor categorization of [ɔ]/[ɒ]$_{O}$ could be explained in a similar manner – in fact, one participant provided post-experiment commentary that indicated that she was familiar with what the dialect variant of this vowel should sound like, but also that she thought the talker's pronunciation did not quite reach this expectation. Thus, it is possible that some listeners' expectations and metalinguistic awareness about dialect speech are based on more extreme pronunciations (likely associated with villages), such that the less extreme pronunciations of a talker from Gavaṛ did not activate their metalinguistic knowledge.

Chapter 6 Comparison of Two Perception Tasks and Comparison of Perception and Production

6.1 Chapter summary

This chapter investigates how perception differs across the two experimental tasks and how production of the relevant vowels relates to perception. After summarizing goals and expectations related to the examination of these questions, the chapter reviews and compares the relevant findings of the analyses that have been presented thus far in other chapters. Specifically, the main findings of the AX and dialect categorization tasks (presented in Chapters 4 and 5 respectively) are summarized and compared in order to address the question of how perception differs across the two tasks and across the two vowel pairs. As the statistical analyses presented in Chapters 4 and 5 (predicting outcomes of the AX and dialect categorization tasks) also incorporated some comparisons of individual-level performance across the two perception tasks and individual-level perception-production relationships, those results are also summarized. The link between production and perception is also explored by comparing which demographic factors were relevant in predicting production and in predicting perception in the two tasks. After this summary of relevant findings from analyses presented in previous chapters, two additional analyses are presented. The first is an individual-level comparison, which explores how each participant's responses to the same stimuli differ across the two perception tasks. The second compares participants' perception and production with criteria for discriminability and/or mergedness used in previous studies, while also comparing participants' productions with their own perception and with the perception of the most accurate discriminators in the sample. This second analysis thus not only explores individual perception-production relationships, but also explores the implications of individual and community-level production-perception relationships for defining vowel merger.

The results of the two new analyses are overall consistent with the results presented in Chapters 4 and 5, in that they suggest that perceptual outcomes are affected by task type, and that discrimination of [a]/[ɒ]$_A$ was unexpectedly poor given the large acoustic differences between

the members of this pair not only in the stimuli, but also in participants' own pronunciations. The results also suggest that the size of vowel differences that are discriminable is highly context dependent, which in turn suggests that local perception of vowel phenomena should be incorporated into context-dependent definitions of merger.

6.2 Goals and expectations

6.2.1 Comparison of perception in AX and dialect categorization tasks

One of the goals of the perception experiment was to investigate potential relationships between participants' performance on the two tasks. This includes not only broad comparison of the results of the two perception tasks, with consideration for how this comparison differs for the two vowel pairs given their differing characteristics, but also investigation of relationships between concrete aspects of individual performance on the two tasks. For example, metrics such as more accurate discrimination or discrimination of smaller intervals on the AX task could relate to categorizing the stimuli in a more expected manner in the dialect categorization task. On the other hand, it could turn out that accurate discrimination is not predictive of performance on the categorization task, such that those with poor discrimination could show accurate categorization, and vice versa. As the two perception tasks included many identical stimuli, this question also investigates how participants' responses to those specific stimuli vary across the two tasks. For example, might participants group stimuli separately in the second task that they were not able to discriminate in the first task? How do their responses to specific stimuli compare across the two tasks? For example, how does their perceptual boundary location in the dialect categorization task compare with their responses to various intervals in the AX task?

Though it is difficult to form specific hypotheses about how the results of the two perception tasks will compare, the literature reviewed in Chapter 3 suggests that a wide variety of social factors can affect perception. Thus, while it is possible that correspondences between responses to the two tasks could be found, it also seems likely that the metalinguistic social prompting of the categorization task will result in some sort of difference between the results of the two tasks. For example, participants could discriminate certain stimuli in the AX task but then group them together (as dialect or non-dialect pronunciations) in the categorization task. This could indicate that listeners are sensitive to small acoustic differences that they do not or cannot necessarily use in categorizing sounds by dialect association, at least given the

categorization options available in the present study. On the other hand, participants could be more successful at identifying stimuli as different once the task is framed through the lens of recognition of dialect association, rather than when they are simply asked to discriminate vowels. In this case, they would group tokens separately in the second task that they had not discriminated in the first task. This would perhaps indicate – at least in the case of [a]/[ɒ]$_{A}$ – that perception is inhibited by the allophonic and/or coarticulatory nature of the vowel differences in the discrimination task more so than in the categorization task, in accordance with the different framing of the tasks. Pilot data suggested that both types of mismatches may be possible, even within the same individual for different lexical items representing the same vowel pair. Thus, it is expected that the results of the comparison of the two perception tasks will be complex.

Given the differences between the vowel pairs found in the results of both perception tasks (see Chapters 4 and 5), it seems likely that characteristics of the vowel pairs themselves could influence the relationships between perception across the two tasks. Specifically, because [a]/[ɒ]$_{A}$ is more socially salient and is also allophonic, it is hypothesized that task type could have a more pronounced influence on its perception than on the perception of [ɔ]/[ɒ]$_{O}$.

6.2.2 Comparison of perception and production

Another broad goal of the perception experiment was to compare performance on the two perception tasks with participants' productions of the relevant vowels. Both across participants and individually, how will the results of the two perception tasks compare with the production of vowel differences? Considering that previous investigations of production and perception of vowel mergers and other vowel phenomena (e.g., Freeman 2023; Fridland & Kendall 2012; Hay et al. 2006; Hay et al. 2013; Kendall & Fridland 2012; Kendall & Fridland 2017; Kettig & Winter 2017; Loakes et al. 2014; Perkell et al. 2004; Warburton 2020; Zhang et al. 2022) have found varying degrees and types of correlation (including no correlation) between the production and perception of individuals or of particular demographic groups, how might variation in perception relate to variation in production? Furthermore, what can the comparison of individual and overall perception and production data tell us about how to define vowel distinction and merger in both perception and production? What are the implications of individual variation for such definitions? For a vowel pair like [a]/[ɒ]$_{A}$, where even those participants who do not use the dialect's enhanced allophony might show small, consistently measurable differences in

production (see Chapter 3), can the perception results inform us about what kinds of differences might be imperceptible or not socially meaningful despite being measurable in production?

As noted in Chapter 3, the potential coarticulatory origin of the [a]/[ɒ]$_A$ allophony could mean that a "merger" of these vowels would never be expected to reach completion, or that small differences between the vowels might not be perceived (at least under some conditions) due to perceptual compensation for coarticulation. It is thus hypothesized that some differences produced by participants and deemed as unmerged and/or discriminable by the threshold method of Stanley and Sneller (2023) and/or other methods and criteria will be smaller than what is generally discriminated in the AX task, which would call into question the appropriateness of such methods/criteria. At the individual level, it is hypothesized that some individuals will produce measurable (i.e., by various criteria unmerged and/or discriminable) differences that are smaller than the smallest differences that they themselves or any of the other participants discriminate in the AX task. Such a finding in the case of [a]/[ɒ]$_A$ would suggest that participants are either not discriminating measurable differences due to their being allophonic, or are interpreting these differences as similar to coarticulatory effects and perceptually compensating for them. Additionally, based on previous findings about links between perception and production (see Section 3.4.1), in which perceptual accuracy or acuity sometimes correlates with aspects of speech production, it is possible that speakers who produce larger differences could show different perceptual patterns from those who produce smaller differences. Nevertheless, as many studies have also found no link between perception and production, it is also possible that no relationship will be found.

As in the case of the other questions explored in the perception experiment, the two vowel pairs under consideration ([a]/[ɒ]$_A$ and [ɔ]/[ɒ]$_O$) could pattern differently with respect to production-perception relationships. In particular, the difference in social salience could lead to differences in perception-production relationships. For example, the salience of [a]/[ɒ]$_A$ could mean that everyone discriminates it easily regardless of their own productions, while a more clear correlation between perception and production might be observed for the less salient [ɔ]/[ɒ]$_O$. On the other hand, differences in phonological status could also influence the relationship between perception and production. For example, [ɔ]/[ɒ]$_O$ could be more easily discriminated by all regardless of production due to its contrastive status.

6.3 Summary and comparison of results of relevant analyses presented in previous chapters

This section provides a synthesis of findings presented separately for each perception task in Chapters 4 and 5, addressing the research goal outlined in Section 6.2.1 by commenting on broad ways in which perception of the two vowel pairs differed across the two perception tasks. This section also begins to address the goal presented in Section 6.2.2 by summarizing which demographic factors were found to be relevant for production of various vowels (see Chapter 2) vs. for their perception in the two different tasks. Though the analyses in Chapters 4 and 5 focused primarily on predicting responses and metrics of performances for the two perception tasks separately, these analyses also included some investigation of individual-level relationships between performance on the two perception tasks and between perception and production. The section below thus summarizes the relevant findings of these analyses, setting the stage for the additional qualitative analyses of individual-level relationships presented in this chapter. Overall, comparison of the two perception tasks reveals that the results for the two vowel pairs differ in ways that can be explained by their phonological and social characteristics. The demographic factors that were relevant for production and for perception in the two tasks also differed and overlapped in interesting ways. On the other hand, results from the quantitative analysis of individual-level production-perception links and of relationships between perception in the two tasks were limited and inconclusive.

6.3.1 General comparison of the two vowel pairs in results of AX and dialect categorization tasks

6.3.1.1 Summary of results of AX task – allophonic vowels discriminated poorly

The analysis of the AX task presented in Chapter 4 showed that discrimination of the two vowel pairs ([a]/[ɒ]$_A$ and [ɔ]/[ɒ]$_O$) deviated from expectations in several ways. First, despite the acoustic differences in the stimuli being considerably larger for [a]/[ɒ]$_A$ than for [ɔ]/[ɒ]$_O$, overall performance was similar across the two vowel pairs based on several metrics (ACCURACY, CONSISTENT INTERVALS ACCURACY, *d'*). The size of the acoustic differences presented in the stimuli was related to performance within but not across vowel pairs; in other words, performance was worse on *pat* and *tsots* (words with smaller acoustic differences) than on *mas* and *pos*. Thus, the size of differences presented was not irrelevant for participants'

discrimination of the stimuli, but it cannot explain why overall performance for [a]/[ɒ]$_A$ was not better than for [ɔ]/[ɒ]$_O$.

The results in relation to both subsets of AX stimuli (consistent and varying intervals) indicated that discrimination of [a]/[ɒ]$_A$ was particularly poor at the end of the continuum that represented small differences from the SEA-like pronunciation [a] (i.e., small differences from step 7). In general, participants had much lower discrimination of the 2-step interval 5-7 than of 2-step intervals elsewhere along the [a]/[ɒ]$_A$ continuum. Among the varying size intervals, no participant's SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED was 5-7 for either word representing the [a]/[ɒ]$_A$ continuum. In contrast, there were participants who discriminated this interval for both words representing the [ɔ]/[ɒ]$_O$ continuum. Like the results in relation to other metrics of performance on the AX task, this result is surprising because of the smaller acoustic differences presented for [ɔ]/[ɒ]$_O$.

As discussed in Chapter 4, these unexpected results can likely be explained by the phonological characteristics of the two vowel pairs. The [a]/[ɒ]$_A$ pair is completely allophonic, while [ɔ]/[ɒ]$_O$ is contrastive despite showing remnants of phonological conditioning in some subsets of vocabulary. Thus, worse discrimination of [a]/[ɒ]$_A$ could be explained by the general tendency for listeners to have difficulty discriminating non-contrastive differences (see Chapter 3 for review of literature that identifies this tendency). As noted in Chapter 4, the particularly poor discrimination toward the SEA-like end of the [a]/[ɒ]$_A$ continuum has several potential explanations. First, since small differences toward this end of the continuum are most analogous to the small, potentially coarticulatory differences identified in Yerevan by Seyfarth and Garellek (2018), it is possible that listeners attributed these differences to the consonantal environment and thus had difficulty discriminating them. Second, since the SEA-like pronunciation [a] is more likely than other variants along the continuum to align with participants' prototypical representation of the phoneme /a/, poor discrimination of small differences from this variant could be explained by the "perceptual magnet effect" (Kuhl 1991).

6.3.1.2 Summary of results of dialect categorization task – social salience facilitates categorization

The analysis of the dialect categorization task presented in Chapter 5 uncovered differences in responses to the two vowel pairs that can also be explained by their differing characteristics. Specifically, responses to the [a]/[ɒ]$_A$ stimuli were more in alignment with the

expected response pattern (i.e., having more "Gavaṛ" responses at the dialect-like end of the continuum and more "not Gavaṛ" responses at the SEA-like end of the continuum) than responses to [ɔ]/[ɒ]$_O$. This tendency emerged from both the analysis of pooled data (using mixed-effects logistic regression models) and analyses of patterns in individual-level metrics like RESPONSE PATTERN and SLOPE of responses. As [a]/[ɒ]$_A$ is likely more salient as a dialect feature than [ɔ]/[ɒ]$_O$, it is possible that social salience played a key role in helping participants to categorize the stimuli in the expected manner.

6.3.1.3 Comparison of overall results of two perception tasks – poor AX discrimination does not hinder dialect categorization

Based on overall comparison of the results of the two perception tasks, it seems that accuracy in AX discrimination is not necessarily related to having the expected response pattern in the dialect categorization task. The discrepancy between the results of the two perception tasks can be seen at the level of both vowel pairs and words. Not only was categorization more as expected for [a]/[ɒ]$_A$ despite similar AX discrimination for the two vowel pairs, but categorization was also more as expected for the word *pat* than for the word *mas*, despite *pat* having been discriminated less accurately in the AX task (likely due to smaller acoustic differences in stimuli). Thus, while it seems that phonological characteristics of the two vowel pairs (specifically, the allophonic nature of [a]/[ɒ]$_A$) are likely responsible for the unexpectedly poor AX discrimination of [a]/[ɒ]$_A$, the social salience of [a]/[ɒ]$_A$ and the social framing of the dialect categorization task may have allowed participants to attend to differences that they did not reliably discriminate in the AX task, helping them to categorize the [a]/[ɒ]$_A$ stimuli in a more expected manner. In other words, the results of the two perception tasks demonstrate how both social factors (e.g., social salience) and linguistic factors (e.g., a sound's phonological status) can influence perception, highlighting how social factors can facilitate perception of allophonic and/or coarticulatory differences (see Section 6.4.1.2 below for examples of responses of individual participants who responded in the expected manner in the dialect categorization task despite poor AX discrimination).

The analysis of perceptual boundary locations presented in Chapter 5 sheds light on some more specific aspects of relationships between the results of the two perception tasks. On average, boundary locations for [a]/[ɒ]$_A$ were not close to the SEA-like end of the continuum (while boundaries for [ɔ]/[ɒ]$_O$ were closer to that end of the continuum). Similarly, in the

consistent intervals portion of the AX task, discrimination was very low toward the SEA-like end of the continuum for [a]/[ɒ]$_A$, but was slightly or substantially higher (depending on the word) toward this end of the continuum for [ɔ]/[ɒ]$_O$. If small differences from [a] (i.e., from the SEA-like endpoint of the continuum) are indeed perceived as the same type of low-level allophonic or coarticulatory phenomena found in Yerevan Armenian, rather than as a manifestation of Gavaṛ's salient [a]/[ɒ]$_A$ allophony, then it makes sense that perceptual boundaries for this vowel in the dialect categorization task would not be near the SEA-like end of the continuum. In other words, for [a]/[ɒ]$_A$, perceptual boundaries in the dialect categorization task were on average not in the area of the continuum that participants had great difficulty discriminating in the AX task, as this area of the continuum might be assumed to be analogous to small coarticulatory differences from [a]. The average boundary location being near the middle of the continuum suggests that it was variants more dramatically different from [a] that were generally perceived as Gavaṛ's characteristic dialectal allophone, while small differences from [a] were likely perceived as the same type of low-level allophonic or coarticulatory phenomena found in Yerevan Armenian. It is nevertheless important to note that the statistical analysis of BOUNDARY LOCATION included participants whose response patterns were in both the expected and opposite directions, meaning that variants more dramatically different from [a] were categorized as non-dialectal by some participants included in the analysis. However, such participants were relatively few for the [a]/[ɒ]$_A$ vowel pair (see **Figure 5.16**), and their boundary locations do not appear to have skewed the data in any substantial way (see Appendix F). In any case, the fact that boundary locations for [a]/[ɒ]$_A$ generally did not fall in areas of poor AX discrimination provides valuable insights about the relationship between discrimination and categorization regardless of the unexpected directionality of some participants' responses. Since boundary locations and areas of relatively high AX accuracy for [ɔ]/[ɒ]$_O$ were both closer to the SEA-like end of the continuum, a general connection between boundary location and relatively accurate AX discrimination can be noted for both vowel pairs.

6.3.2 Findings related to quantitative comparison of the two perception tasks – weak and contradictory effects

The analyses presented in Chapter 5 also included some direct quantitative comparisons of the results of the two perception tasks. Specifically, the linear and logistic regression models predicting individual-level metrics of performance on the dialect categorization task tested

individual-level metrics from the AX task as predictors. The findings of this analysis were limited and difficult to interpret. For [a]/[ɒ]$_A$, there was a sub-significant tendency for participants with higher *d'* values to have steeper slopes (whether in the expected or opposite direction, i.e., ABSOLUTE VALUE OF SLOPE). In the models predicting RESPONSE PATTERN as a categorical variable, effects of *d'* and AX ACCURACY were found to be opposite in terms of the trends found in relation to each vowel pair for each of the two metrics (see **Figures 6.1** and **6.2** below; these and all other data visualizations in this chapter were created in R [R Core Team 2021] using *ggplot2* [Wickham 2016]). Nevertheless, the only one of these trends to have a slope significantly different from zero (p=0.038) was the effect of *d'* in relation to RESPONSE PATTERN of [a]/[ɒ]$_A$, according to which participants with higher *d'* values had more of the expected response pattern. Thus, while there were some weak indications that accurate discrimination in the AX task might relate to categorizing stimuli more categorically and/or in a more expected manner, these effects were limited to only a few metrics (both in terms of the predictors and outcome variables), were not always significant, and were sometimes contradictory.

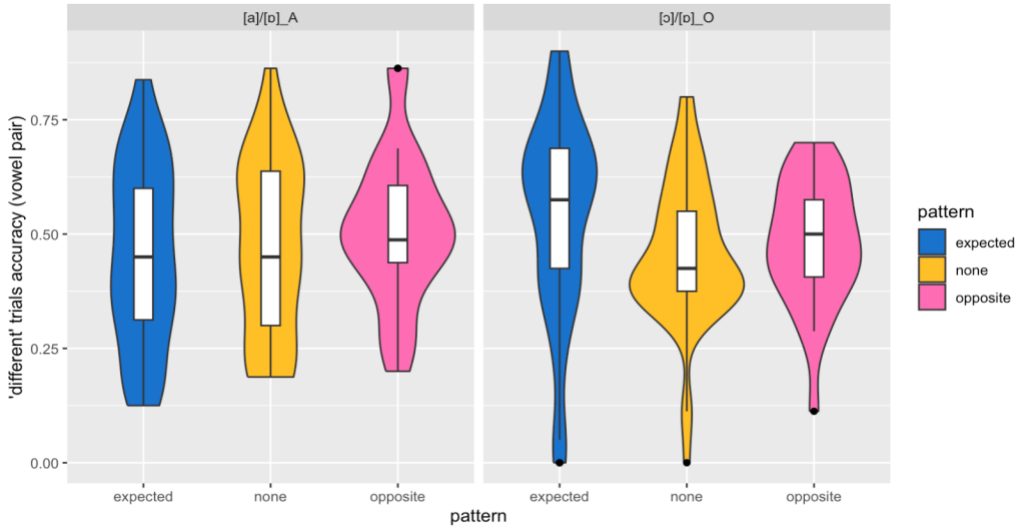


Figure 6.1: Violin plots showing the relationship between ACCURACY in the AX task (calculated for each vowel pair) and RESPONSE PATTERN in the dialect categorization task (also pooled for each vowel pair). The plots exclude Participant 36, who was excluded from all analyses involving the AX task. The box represents the range between the first and third quartiles (interquartile range), while the horizontal line represents the median. The whiskers represent data points that lie up to 1.5 times the interquartile range from the edge of the box, while black dots represent outliers.

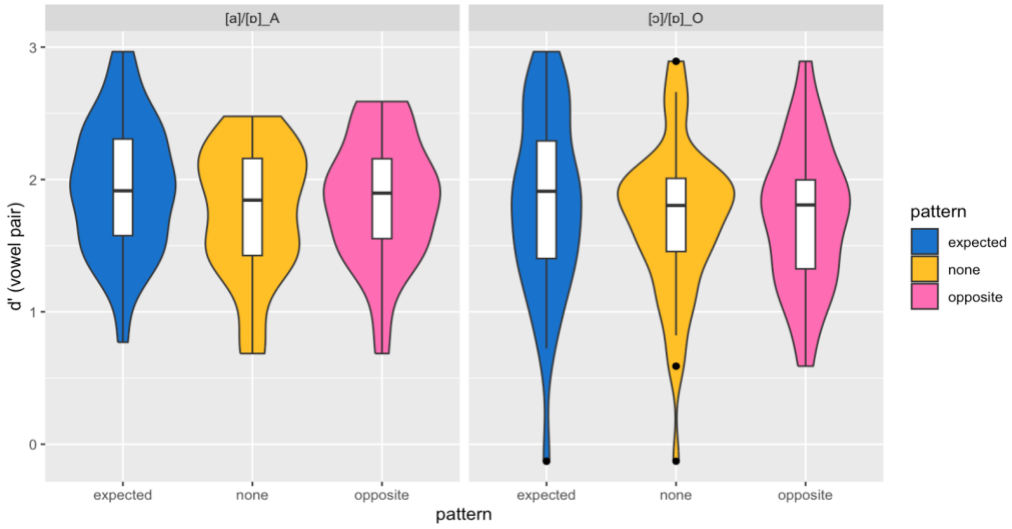


Figure 6.2: Violin plots showing the relationship between *d'* in the AX task (calculated from pooled data for each vowel pair) and RESPONSE PATTERN in the dialect categorization task (also pooled for each vowel pair). The plots exclude Participant 36, who was excluded from all analyses involving the AX task. The box represents the range between the first and third quartiles (interquartile range), while the horizontal line represents the median. The whiskers represent data points that lie up to 1.5 times the interquartile range from the edge of the box, while black dots represent outliers.

6.3.3 Comparison of production and perception

6.3.3.1 Comparison of demographic factors relevant for production and perception

The analyses of production (Chapter 2), the AX task (Chapter 4), and the dialect categorization task (Chapter 5) found differing but somewhat overlapping results in terms of the demographic factors that predicted production and perception. As in the 2022 production study, GENDER was the demographic characteristic that most consistently predicted production, with men having more dialect-like productions of some vowels. In contrast, the results of the AX task showed differences in performance based on BIRTH YEAR and EDUCATION LEVEL, with older people and people without higher education discriminating the stimuli more successfully (see **Figures 6.4-6.5** below for examples of these trends). Analysis of the dialect categorization task also suggested that older people categorized stimuli more accurately (i.e., in a more expected manner) than younger people (see **Figures 6.6-6.7** below), but also suggested that women categorized in a more expected manner than men (see **Figure 6.8** below). Thus, the results of the perception study suggest that there are demographic factors (BIRTH YEAR, EDUCATION LEVEL) that

are relevant for (various types of) perception, but not for production. The finding related to the role of GENDER in the dialect categorization task is interesting in that it suggests that this predictor relates to the production and perception of dialect vowels in opposite ways. In other words, while men use some dialect vowels more than women in production, they seem to be worse at recognizing those vowels as dialectal variants in perception. Overall, comparison of the production and perception results thus suggest that the social factors that are relevant for production and perception are different, and that even a factor that is relevant to both production and perception can have differing relationships with them.

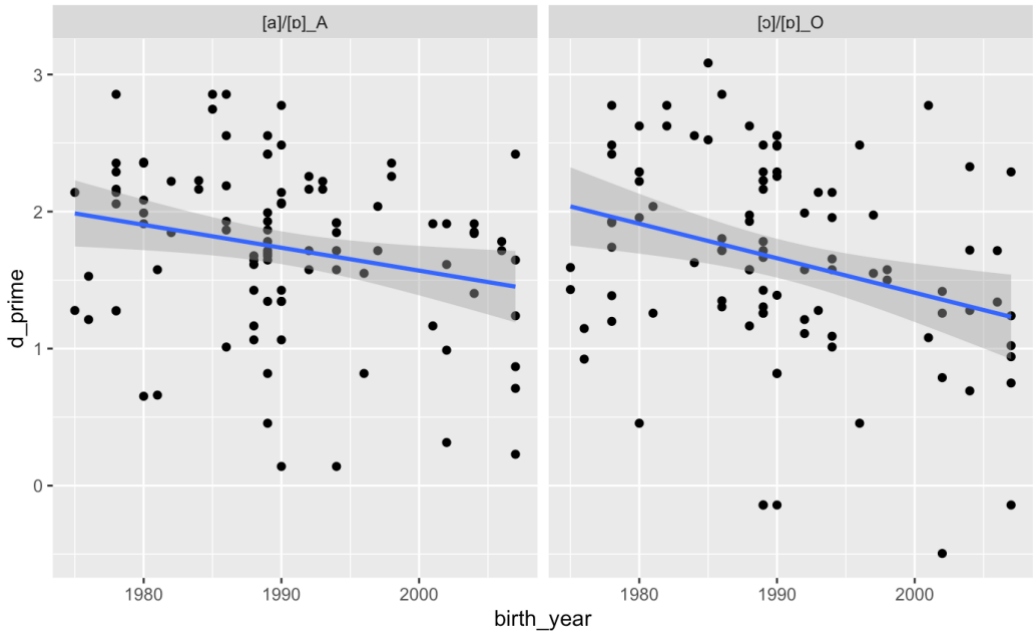


Figure 6.3: Relationship between BIRTH YEAR and *d'* in the AX task. Each dot represents one participant's *d'* value for one individual word. The *d'* values are grouped by vowel pair. The overlaid regression lines have a negative slope for both vowel pairs, indicating that younger people had lower *d'* values (i.e., lower accuracy on the AX task).

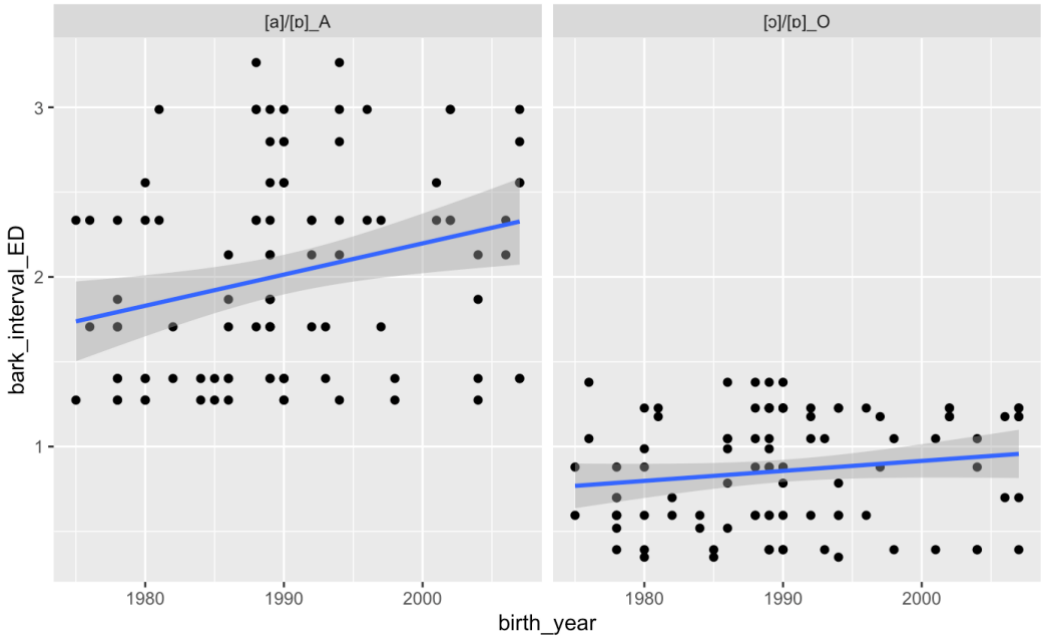


Figure 6.4: Relationship between BIRTH YEAR and SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED from the AX task (in Bark). Each dot represents one participant's interval value for one individual word. The values are grouped by vowel pair. The slopes of the overlaid regression lines indicate that the general age effect (younger people having larger intervals [i.e., worse discrimination]) found in pooled models in Chapter 4 was driven by [a]/[ɒ]$_A$ more so than by [ɔ]/[ɒ]$_O$, as the former has a more strongly positive slope.

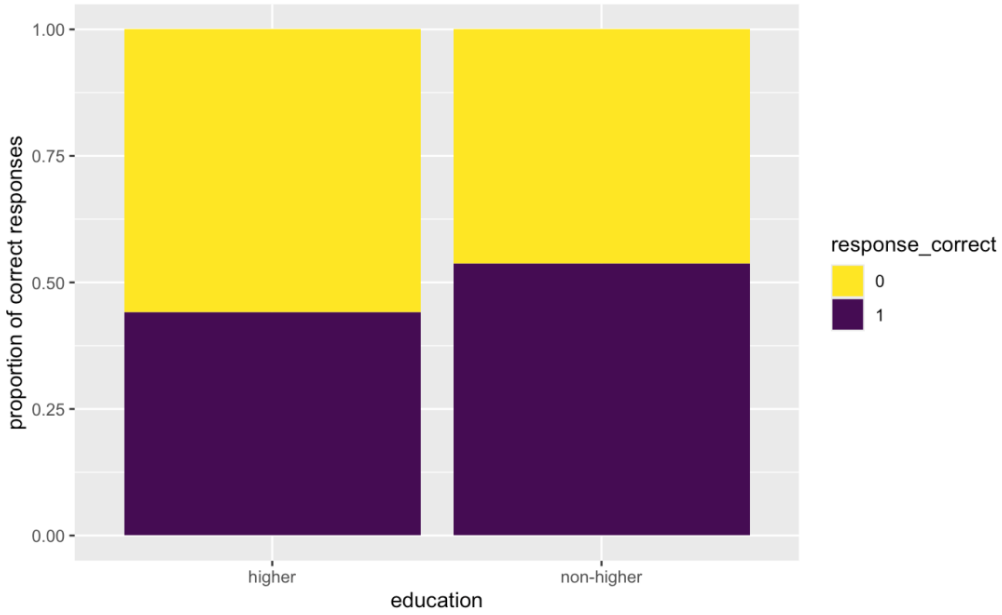


Figure 6.5: Proportion of correct responses by EDUCATION LEVEL for AX varying intervals for [a]/[ɒ]$_A$. Purple represents correct responses, while yellow represents incorrect responses (i.e. failure to discriminate the interval).

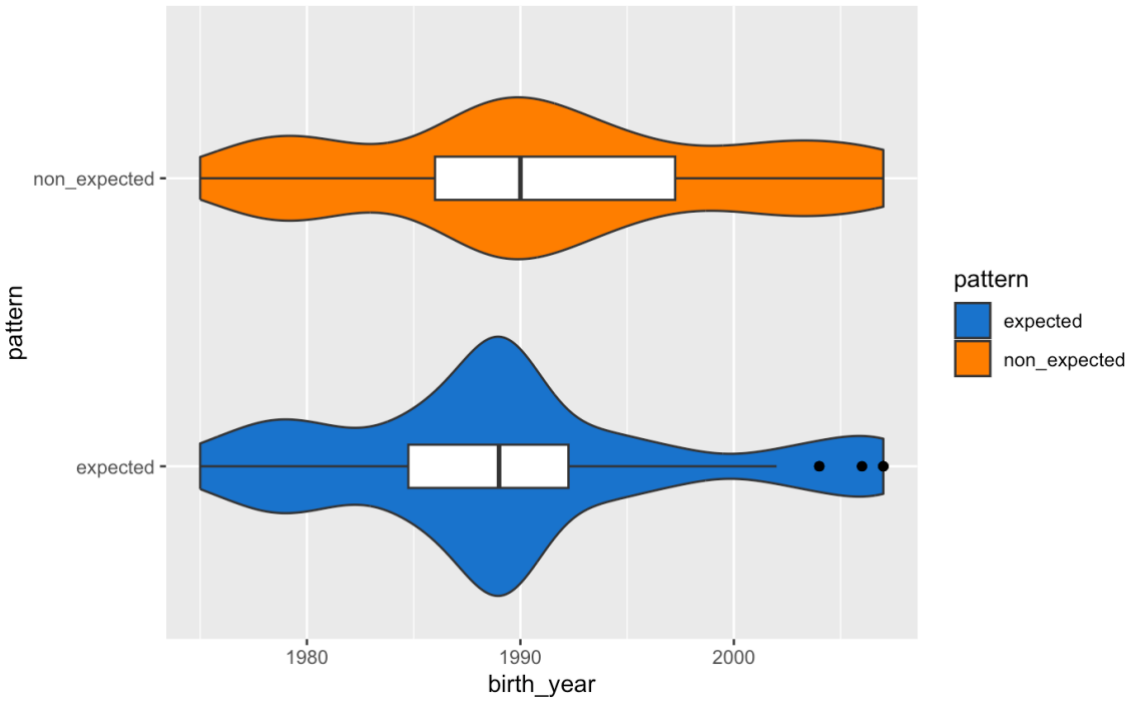


Figure 6.6: Violin plots showing the relationship between BIRTH YEAR and RESPONSE PATTERN (expected vs. non-expected) in the dialect categorization task. The box represents the range between the first and third quartiles (interquartile range), while the horizontal line represents the median. The whiskers represent data points that lie up to 1.5 times the interquartile range from the edge of the box, while black dots represent outliers. The effect of BIRTH YEAR on RESPONSE PATTERN was sub-significant (p=0.079) in models that pooled both vowel pairs without interactions (presented in Chapter 5), such that older people had more expected response patterns.

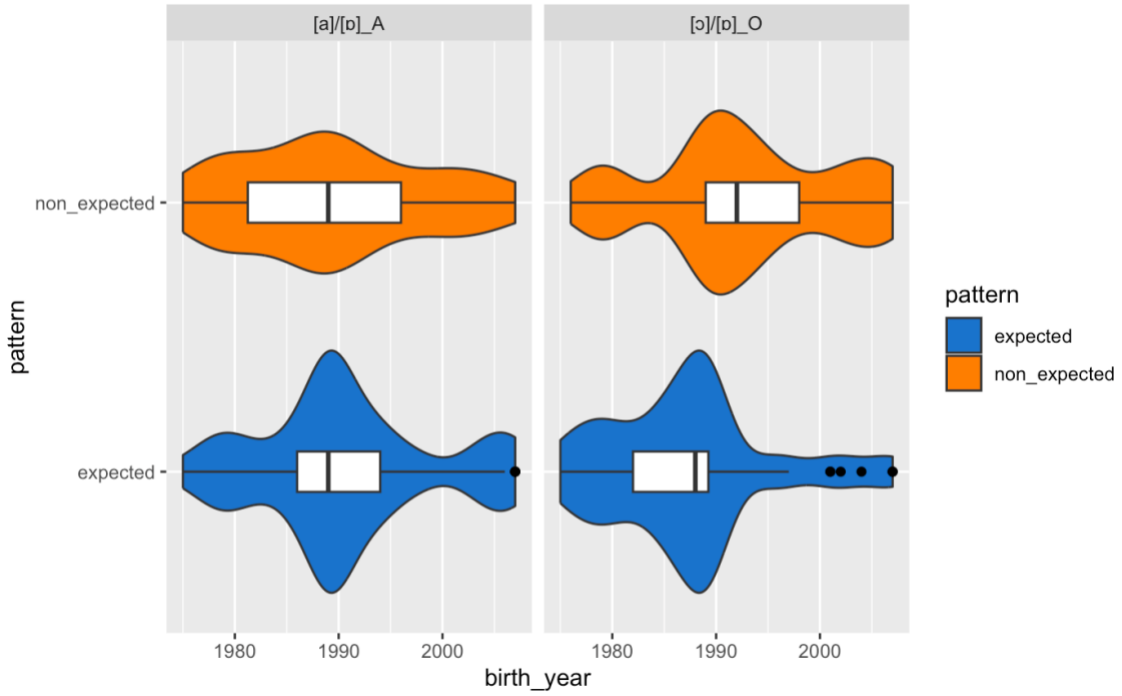


Figure 6.7: Violin plots showing the relationship between BIRTH YEAR and RESPONSE PATTERN (expected vs. non-expected) in the dialect categorization task separately for each vowel pair. The box represents the range between the first and third quartiles (interquartile range), while the horizontal line represents the median. The whiskers represent data points that lie up to 1.5 times the interquartile range from the edge of the box, while black dots represent outliers. The plots indicate that the weak general effect of BIRTH YEAR shown in **Figure 6.6** above is driven primarily by [ɔ]/[ɒ]$_{O}$.

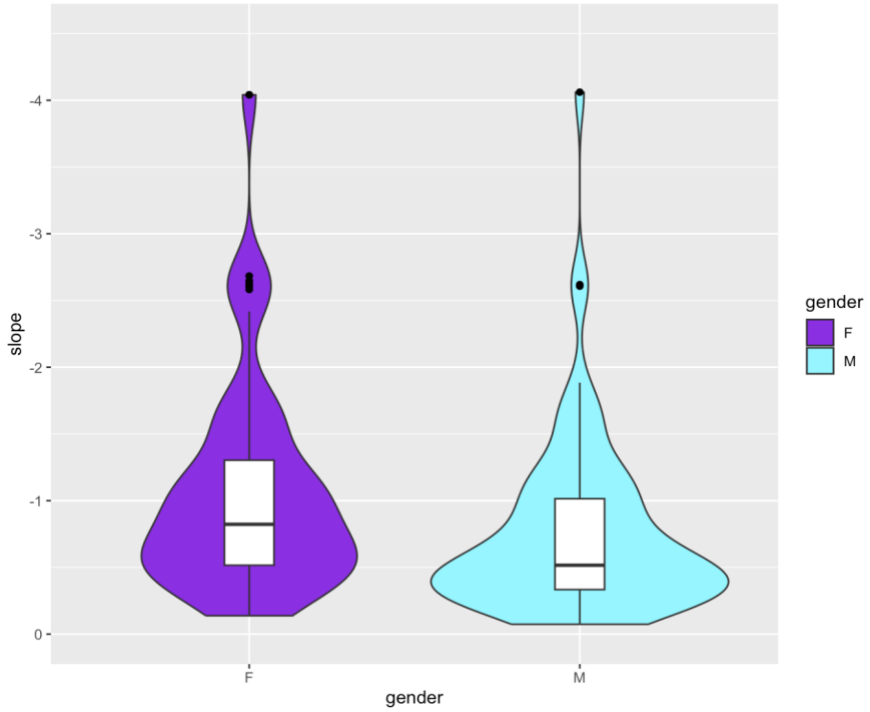


Figure 6.8: Violin plots showing the relationship between GENDER and SLOPE in the dialect categorization task, among only those participants who showed the expected response pattern. The box represents the range between the first and third quartiles (interquartile range), while the horizontal line represents the median. The whiskers represent data points that lie up to 1.5 times the interquartile range from the edge of the box, while black dots represent outliers. The y-axis is inverted, such that higher (more negative) values correspond with steeper slopes. The plots show that women had steeper (more negative) slopes, indicating more categorical response patterns.

6.3.3.2 Quantitative comparison of production and perception – absent, inconclusive or weak trends

Quantitative findings discussed in Chapters 4 and 5 related to links between production and perception were limited. Relationships between production (as measured by Pillai scores and ED-median between the relevant vowel categories) and perception were not found in any analyses of the AX task. As noted in Chapter 5, in the pooled logistic regression models predicting responses to the dialect categorization task, opposite trends were found in relation to the two vowel pairs. Specifically, for [a]/[ɒ]$_{A}$, participants with greater distances between the relevant vowel categories in production (i.e., more dialect-like pronunciations) chose "Gavaṛ"

less (significant trend), while for [ɔ]/[ɒ]$_{O}$, those with less overlap in production (i.e., more dialect-like pronunciations) chose "Gavaṛ" more (sub-significant trend). These trends are difficult to interpret not only because of their opposing directionality, but also because the outcome variable of these models is not indicative of participants' overall response pattern. It is nevertheless interesting that participants with more dialect-like pronunciations of the salient [a]/[ɒ]$_{A}$ pair chose "Gavaṛ" less overall in categorizing this pair, as this could suggest that participants with more extreme dialect-like pronunciations of this pair might also only perceive the most extreme pronunciations of [ɒ]$_{A}$ as dialectal (leading to lower overall "Gavaṛ" responses).

In the models predicting individual-level metrics, there was a sub-significant tendency for participants with more dialect-like productions (greater distances, less overlap between the relevant vowels) to have perception boundaries further toward the dialect-like (left) end of the continuum for [a]/[ɒ]$_{A}$. This result aligns with the aforementioned finding that participants with more dialect-like pronunciations of [a]/[ɒ]$_{A}$ chose "Gavaṛ" less overall. Together, these findings suggest a link between perception and production of [a]/[ɒ]$_{A}$, such that people with more distinct (i.e., dialect-like) vowel categories in production needed a larger difference from the standard variant in order to perceive the dialect variant, thus having boundaries further toward the left and fewer overall "Gavaṛ" responses. These trends were nevertheless weak, and it is important to acknowledge that the analysis of BOUNDARY LOCATION in general relies on a methodology that may not have been ideal for analysis of the given experimental task (see discussion in Section 5.6.1). Furthermore, as the models predicting BOUNDARY LOCATION included participants with both expected and opposite response patterns, the interpretation of the results in relation to this metric is less straightforward for those (relatively few in the case of [a]/[ɒ]$_{A}$) participants whose response pattern did not have the expected directionality (i.e., who had a categorical pattern, but classified the SEA endpoint as "Gavaṛ"). Nevertheless, as the figures in Appendix F show, the tendency for participants with more dialect-like vowels to have perception boundaries further toward the left (dialect-like) end of the continuum was driven primarily by participants with the expected response pattern, providing support for the existence of a production-perception link.

6.4 Additional qualitative analyses

As demonstrated by the summary above, although the analyses presented in Chapters 4 and 5 showed interesting results in terms of overall differences between perception of the two

vowel pairs and broad demographic trends in production and perception, the analyses did not uncover any strong and consistent relationships between individual-level performance on the two perception tasks or between individual perception and production. The findings in relation to these comparisons seem especially weak against the backdrop of findings in relation to differences in responses across the vowel pairs and words, which were more consistent and usually reached statistical significance. Though the analyses presented in Chapters 4 and 5 already incorporated individual-level metrics of performance on the two perception tasks and individual-level production metrics as predictors and outcome variables in statistical models, the sections below present additional individual-level analyses, which examine performance across the two perception tasks and perception-production relationships in more qualitative ways.

6.4.1 Additional comparison of individual-level performance across the two perception tasks

6.4.1.1 Goals and methods

The overarching goal of this qualitative individual-level analysis is to explore the extent to which performance on the AX task (in terms of which intervals each participant discriminated well in comparison with other intervals) could explain aspects of performance on the dialect categorization task (specifically, categoricity of the response pattern and location of the perceptual boundary), as well as the extent to which participants varied in this regard.

The analysis included only those four words that served as stimuli in both perception tasks (*pat, mas, pos, tsots*). For each participant for each word, the steps outlined in **Figure 6.9** were followed in order to categorize the participant by how their responses in the two perception tasks related to one another. The analysis resulted in participants being grouped into various categories, some of which contain further sub-categories. **Figure 6.10** presents an alternative way of visualizing the analysis, showing the resulting groupings and how they relate to one another. The colors used in both **Figure 6.9** and **Figure 6.10** correspond with the visualizations in Section 6.4.1.2, which show the relative amounts of participants in each category or sub-category for each word.

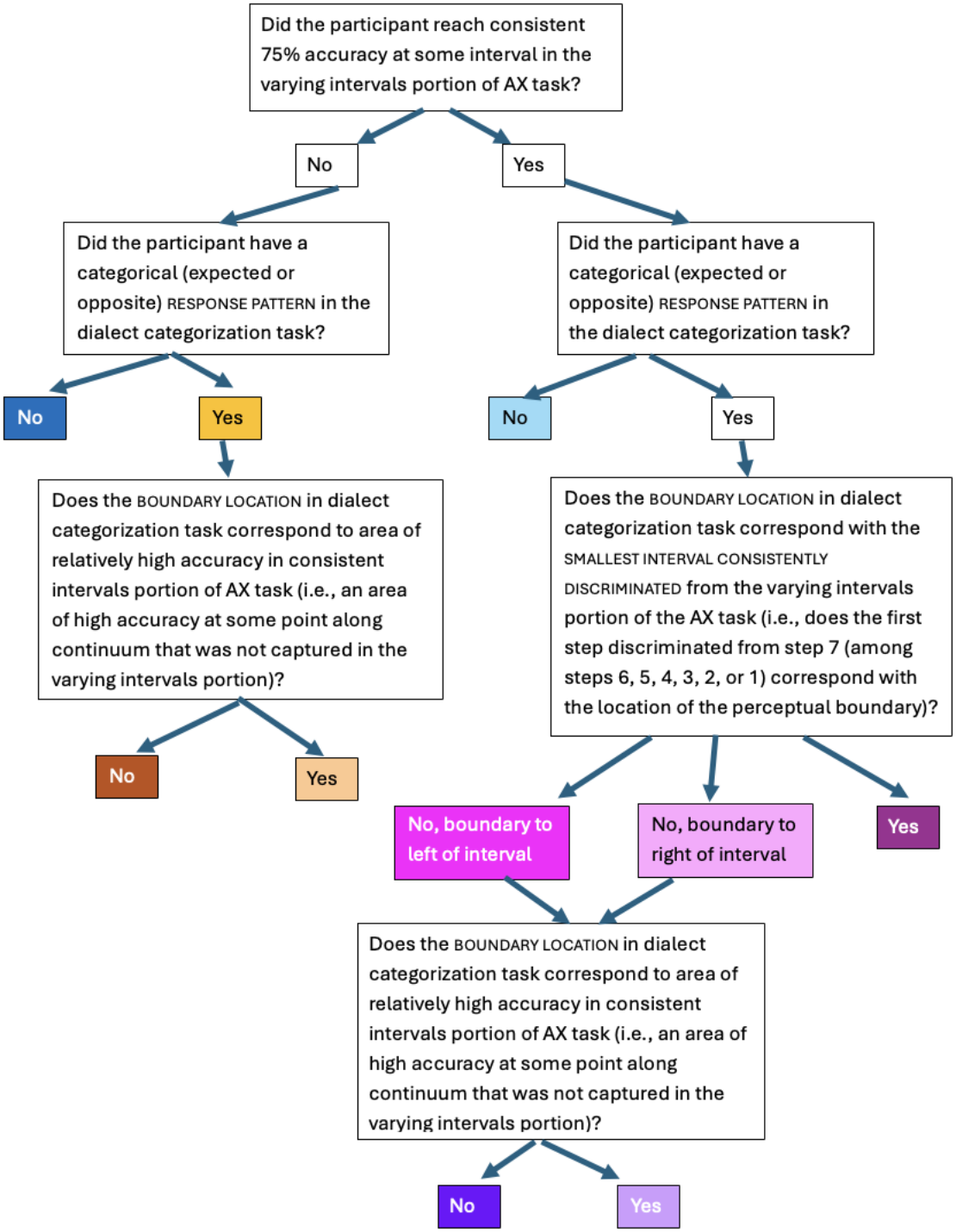


Figure 6.9: Process of categorizing participants based on their responses to the AX and dialect categorization tasks. This process used the RESPONSE PATTERN, SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED and BOUNDARY LOCATION metrics described in Chapters 4 and 5.

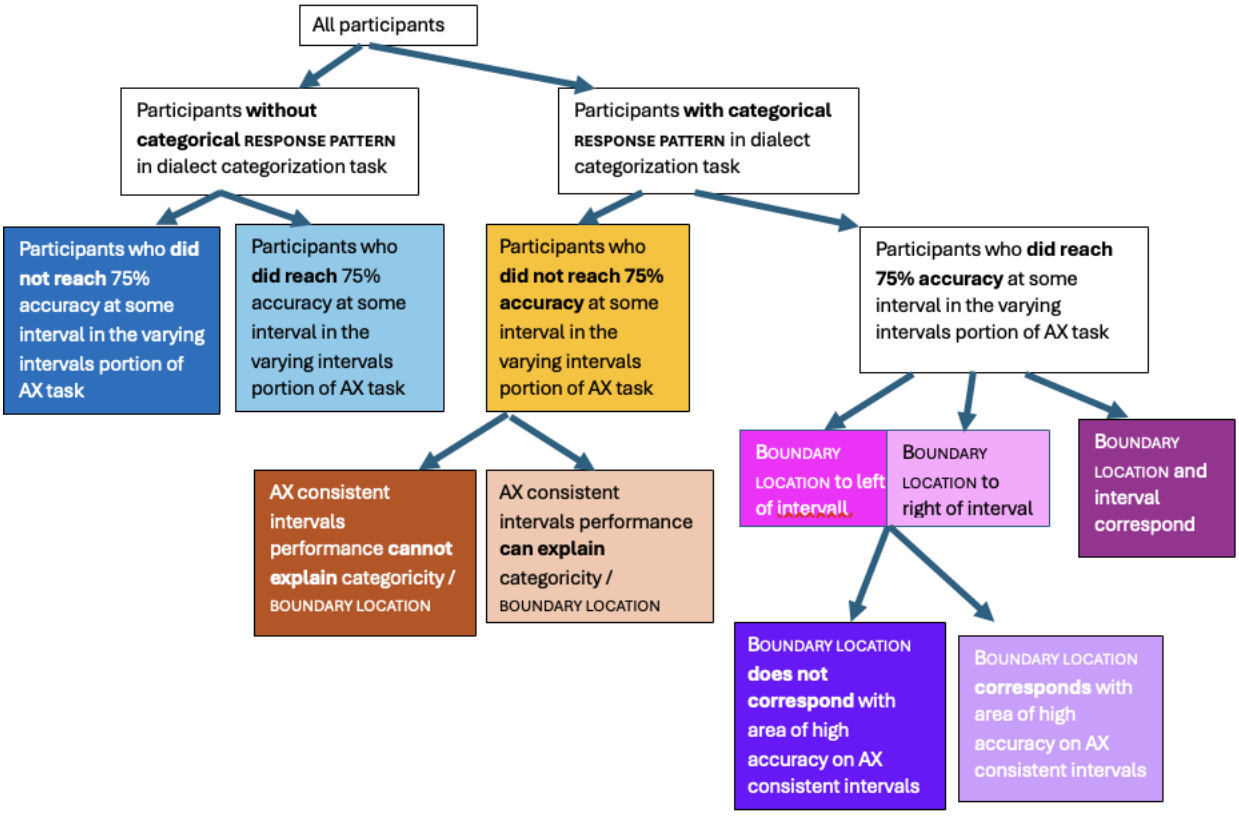


Figure 6.10: Groupings and sub-groupings of participants resulting from the process of categorizing participants based on their responses to the AX and dialect categorization tasks. This process used the RESPONSE PATTERN, SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED and BOUNDARY LOCATION metrics described in Chapters 4 and 5.

Comparison of the dialect categorization BOUNDARY LOCATION with areas of relatively higher vs. lower accuracy on the consistent intervals portion of the AX task was done for all four stimuli words in the case of those participants who never reached consistent 75% accuracy on the varying intervals portion (the results of this comparison are represented by the dark orange and light orange boxes in **Figures 6.9-6.10**). However, for those participants who did reach 75% accuracy (i.e., whose SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED was the interval 1-7 or smaller), this comparison was only done for *pat* and *mas* (the results of this comparison are represented by the lowermost [dark and light violet] boxes in **Figures 6.9-6.10**). This methodological decision was made in connection with the fact that the consistent intervals presented for *pos*/*tsots* consisted of 3 steps rather than 2 steps, such that there was more overlap between the portions of the continuum covered by each interval. Accordingly, in the case of any given participant, it was more difficult to say if the BOUNDARY LOCATION corresponded with an

area of higher accuracy, as a given BOUNDARY LOCATION could be considered to correspond with up to three different intervals (out of four total), which could have very different accuracy levels (for example, a boundary at 3.5 would fall within the intervals 1-4, 2-5, and 3-6). Participants who never reached 75% accuracy on the varying intervals portion also generally had lower accuracy on the consistent intervals portion of the AX task, which made the task of determining if the boundary location corresponded with an area of higher AX accuracy less ambiguous.

Accuracy at a given AX interval from the consistent intervals portion was considered to be relevant for explaining BOUNDARY LOCATION in cases where the boundary location fell within the interval. For example, if a participant had relatively high accuracy at the interval 1-3, and their BOUNDARY LOCATION was 2.5, then their accuracy at the interval 1-3 would be considered relevant for explaining this location. Accuracy at the interval 2-4 would also be considered relevant. In rare cases where a BOUNDARY LOCATION was a perfect whole number (e.g. 6.0), intervals that included this number as their endpoint could be considered relevant. For example, in the case of a boundary at 6.0, accuracy at both the intervals 4-6 and 5-7 would be considered. In cases where a participant's perceptual boundary was not located within the 7-step continuum, their AX performance was considered relevant for explaining the boundary location in cases where they had relatively high accuracy at the interval adjacent to the relevant end of the continuum. For example, if a participant's boundary was at 0, relatively high accuracy at the interval 1-3 would be considered relevant for explaining the boundary.

For the purposes of this analysis, "relatively high accuracy" (as referenced in **Figure 6.9**) at a certain interval in the consistent intervals portion of the AX task was defined as any level of accuracy that was higher than the participant's minimum accuracy in the consistent intervals portion. In other words, 25% accuracy could be considered relatively high if the participant's minimum accuracy was 0, but if the participant's minimum accuracy was 25%, 50% or higher would be considered relatively high. The goal of performing the analysis in this way was to be maximally inclusive of cases where a participant's AX performance might have any connection at all with the location of their perceptual boundary, such that only cases where there was clearly no relationship would be labeled as such. The only exception to this maximally inclusive principle came from Participant 32's responses for the word *tsots*. This participant had a categorical (opposite of expected pattern) response pattern in the categorization task, and their BOUNDARY LOCATION corresponded with an area of higher accuracy from the AX task. However,

that area of relatively high accuracy (25% accuracy compared to the participant's minimum of 0) consisted of a single correct response, which was the participant's only correct response not only in the consistent intervals portion, but in the entire AX task for the given word (i.e., they did not successfully discriminate any larger intervals, consisting of 6, 5, or 4 steps, and only had a single instance of successfully discriminating a 3-step interval). The context of this overall response pattern suggests that the participant's single correct response to a 3-step interval was likely a mistake, and thus should not be considered a meaningful explanation for their categorical response pattern in the second task.

Correspondence of the BOUNDARY LOCATION with the SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED from step 7 in the AX task was defined as the two numbers being within ±0.5 of one another. For example, if a participant's SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED from step 7 was step 3, a perceptual boundary anywhere between 2.5 and 3.5 would be considered to correspond with this. Participants grouped into this category are represented by the reddish purple boxes in **Figures 6.9-6.10**. The goal of comparing these two metrics was to ascertain the degree to which participants' perception of variants as dialectal or non-dialectal (as quantified by their BOUNDARY LOCATION in the dialect categorization task) could be said to be related to discrimination of those variants from the canonical SEA variants in the varying intervals portion of the AX task (step 7). In other words, the analysis sought to determine if there was some relationship between the first step that participants were able to consistently discriminate from step 7 in the AX task and the first step (quantified by BOUNDARY LOCATION) that they perceived as dialectal in the dialect categorization task. The two metrics corresponding would suggest that discrimination of variants from step 7 and perception of those variants as dialectal are somehow related, while the metrics not corresponding would suggest that the two perceptual behaviors are more independent from one another.

The results section below presents visualizations that compare the number of participants belonging to each of the categories listed above across the different stimuli words. Several examples of individual participants' results are also presented, in order to illustrate the different types of correspondences that were found across the two perception tasks.

6.4.1.2 Results and discussion

The results of this analysis show that participants exhibited a wide variety of combinations of response patterns, most of which were found across all four stimuli words.

Figure 6.11 visualizes AX performance (in terms of having reached or not reached consistent 75% accuracy in the varying intervals portion) for those participants who did not have a categorical RESPONSE PATTERN in the dialect categorization task. As this figure shows, accurately discriminating stimuli in the AX task did not guarantee that a participant would show a categorical pattern in the dialect task, as for each of the four words there were participants who did not show a categorical RESPONSE PATTERN despite good AX discrimination (light blue column).

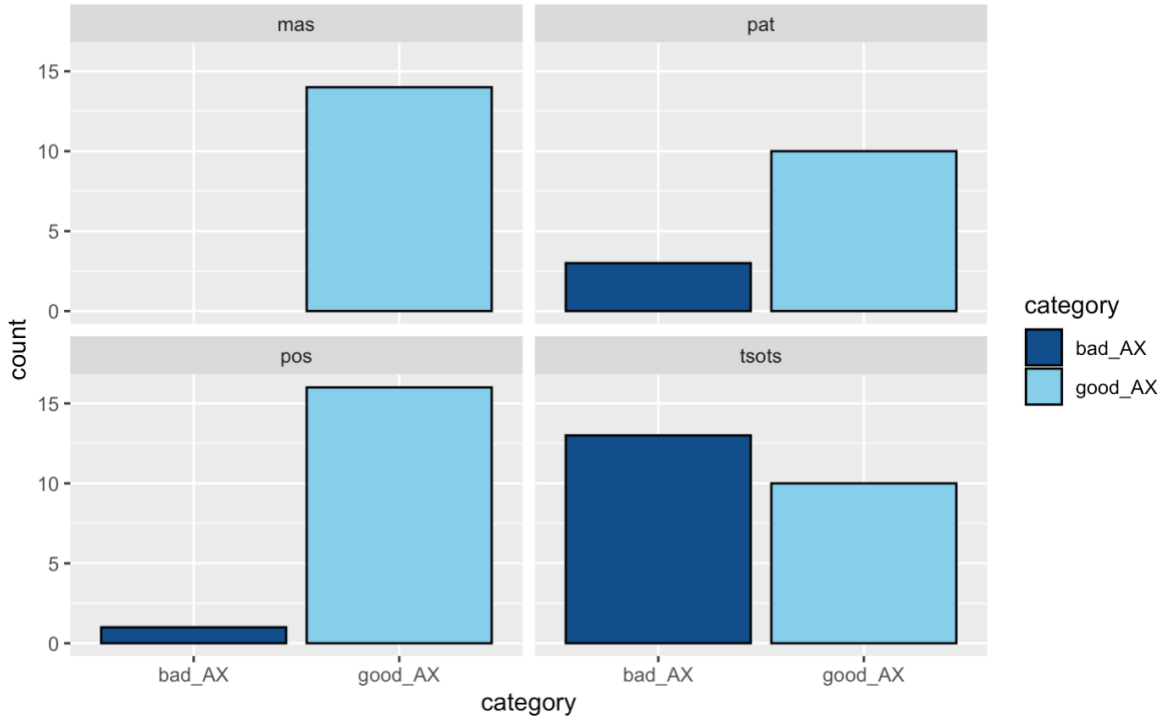


Figure 6.11: Amount of participants who did not show a categorical RESPONSE PATTERN (either in the expected or opposite direction) for each word in the dialect categorization task. Participants in the "good_AX" column reached consistent 75% accuracy for some interval in the varying intervals portion of the AX task, but nevertheless did not have a categorical response for the given word in the dialect categorization task. Participants in the "bad_AX" column did not reach 75% accuracy at any interval in the varying intervals portion of the AX task for the given word.

Figure 6.12 presents a comparison of BOUNDARY LOCATION (for those participants who did show a categorical RESPONSE PATTERN in the dialect categorization task) with the SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED by those participants in the varying intervals portion of the AX task. As the figure shows, BOUNDARY LOCATION had a variety of relationships with AX responses for all four words. Specifically, for all four words, there were participants who had categorical patterns in the dialect categorization task despite never reaching consistent 75%

accuracy on the varying intervals portion of the AX task (yellow column), participants whose BOUNDARY LOCATION corresponded with the SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED from step 7 (reddish purple column), and participants whose boundary was located to the left or right of this interval (bright and pale pink columns – see **Figure 6.17** for further comparison of these participants' BOUNDARY LOCATION with AX results). **Figure 6.13** below provides an example of the individual data of a participant whose BOUNDARY LOCATION and SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED corresponded (reddish purple column).

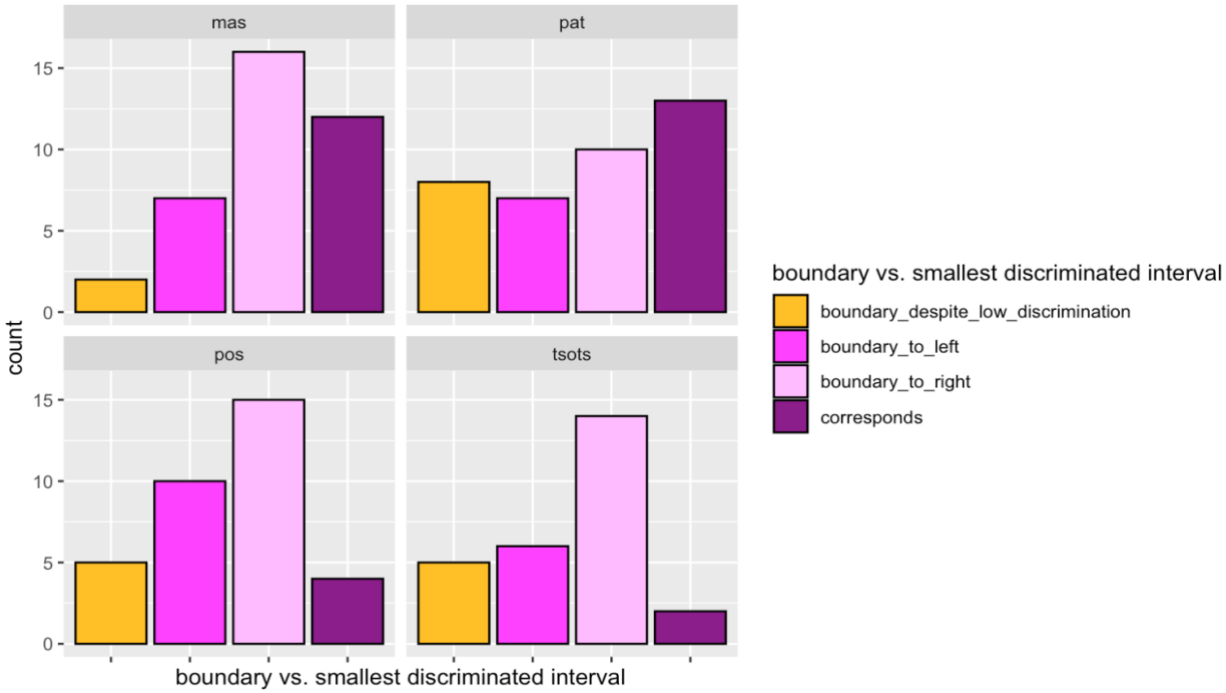


Figure 6.12: Amount of participants for each word who showed a categorical RESPONSE PATTERN (either in the expected or opposite direction) in the dialect categorization task. Participants are sorted according to the relationship between their BOUNDARY LOCATION from the categorization task and their responses to the same word in the AX task. The yellow column represents participants who had a categorical response despite never reaching 75% consistent accuracy in the varying intervals portion of the AX task. The bright and pale pink columns represent participants whose BOUNDARY LOCATION was to the left or right of their SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED from step 7 in the varying intervals portion of the AX task. The purple column represents participants whose BOUNDARY LOCATION and SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED corresponded.

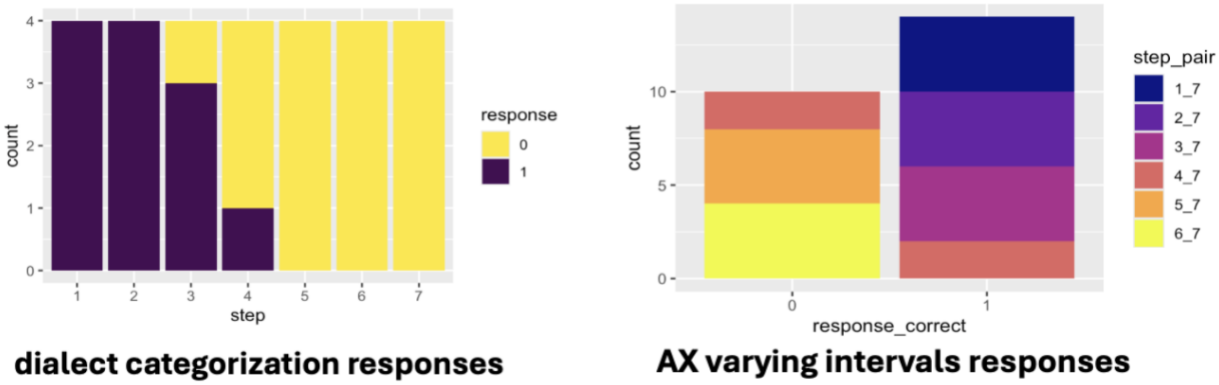


Figure 6.13: Example of participant (Participant 13; *pat*) whose **BOUNDARY LOCATION corresponded with the SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED from step 7 in the AX varying intervals portion**. The SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED was 3-7, and the participant's BOUNDARY LOCATION was 3.5. The left graph is the participant's responses to the dialect categorization task (purple (1) = "Gavaṛ"; yellow (0) = "not Gavaṛ"). The graph on the right represents the participant's responses to AX varying intervals (right column = correct response; left column = incorrect response; colors represent different intervals).

Figure 6.14 summarizes whether the BOUNDARY LOCATION was in an area of relatively high accuracy on the consistent intervals portion of the AX task for those participants who did not reach consistent 75% accuracy at any interval on the varying intervals portion of that task. As the figure shows, the locations of boundaries (and, consequently, the categorical nature of participants' responses) were not always explainable by participants' performance on the AX consistent intervals. For some words, such participants (dark orange column) are more numerous than those for whom AX performance does explain the boundary location (light orange column). **Figures 6.15** and **6.16** provide examples of the responses of individual participants representing the dark orange and light orange columns respectively.

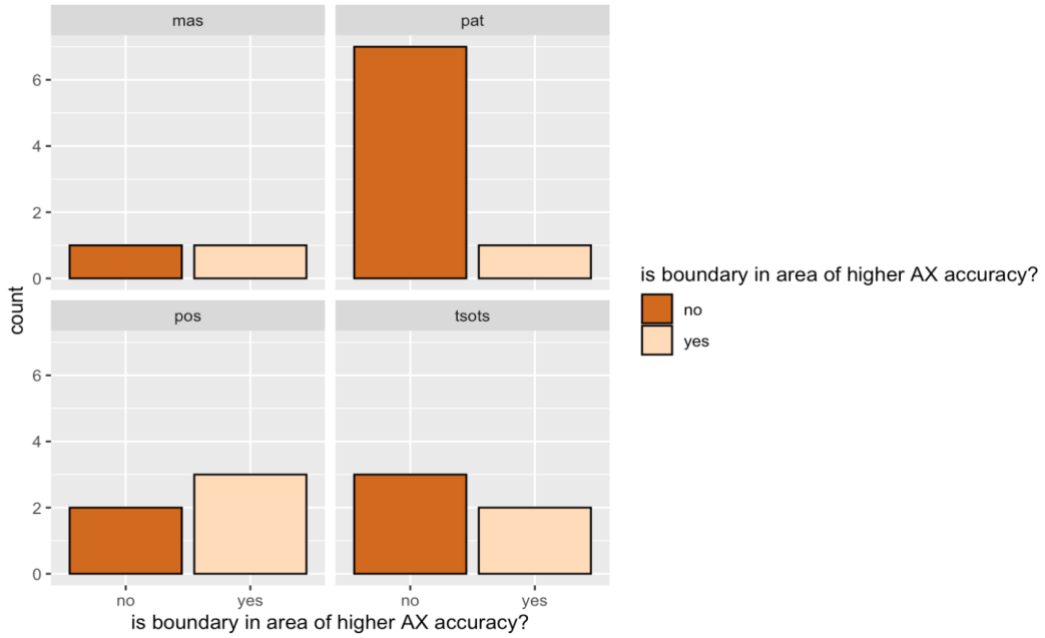


Figure 6.14: Comparison of BOUNDARY LOCATION with areas of relatively high accuracy on the consistent intervals portion of the AX task (among those participants who **never reached** consistent 75% accuracy on the AX varying intervals portion for the relevant word).

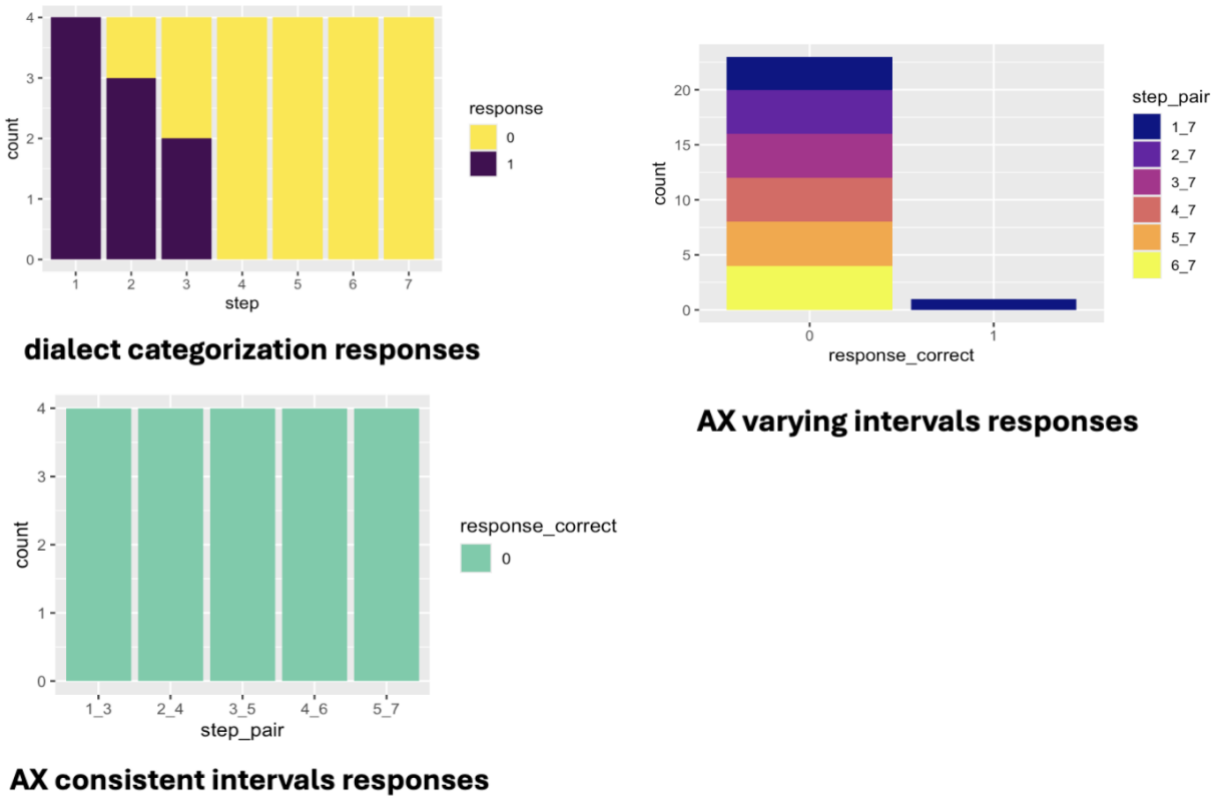


Figure 6.15: Example of participant (Participant 24; *pat*) who **had a perceptual boundary in the dialect categorization task despite never reaching consistent 75% accuracy in the varying intervals portion of the AX task. The BOUNDARY LOCATION is also not explainable by AX accuracy elsewhere along the continuum** (the participant had 0% accuracy for all 2-step intervals). The upper-left graph is the participant's responses to the dialect categorization task (purple (1) = "Gavaṛ"; yellow (0) = "not Gavaṛ"). The lower-left graph represents the participant's accuracy on AX consistent intervals (dark blue=correct; green=incorrect). The graph on the right represents the participant's responses to AX varying intervals (right column = correct response; left column = incorrect response; colors represent different intervals). The participant's BOUNDARY LOCATION was 2.75.

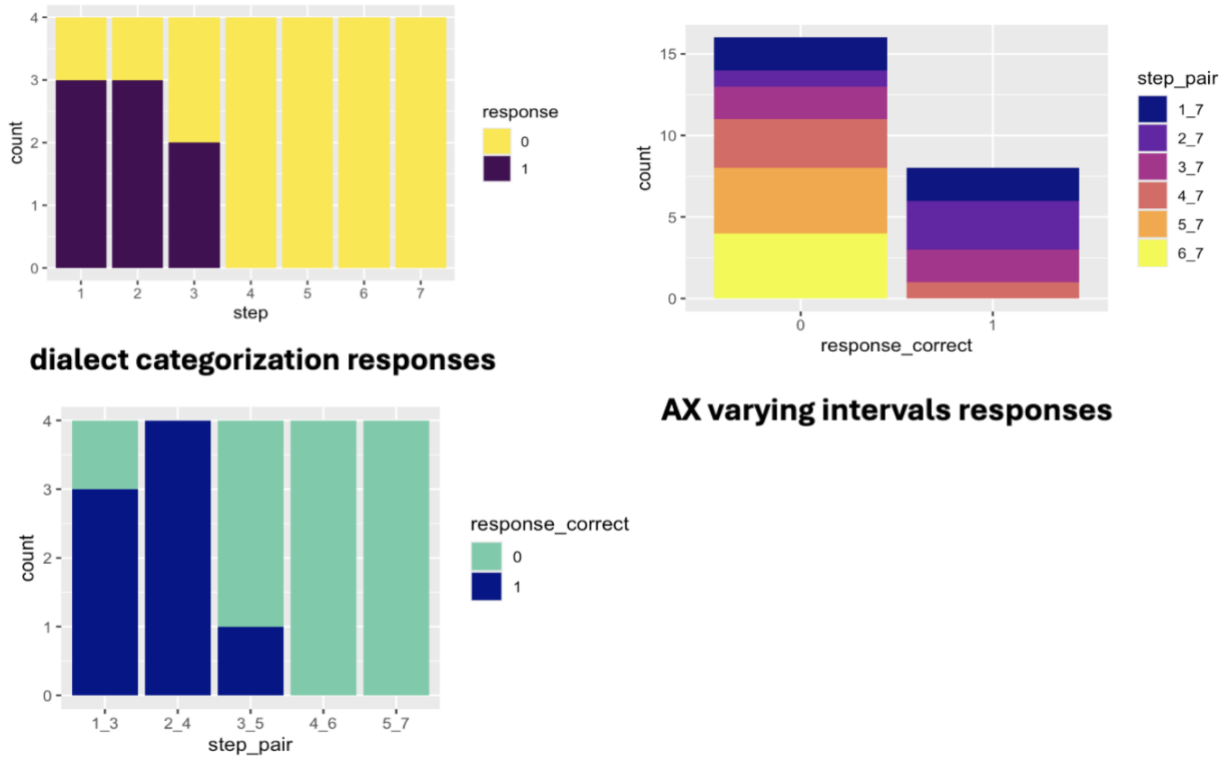


Figure 6.16: Example of participant (Participant 20; *mas*) who **had a perceptual boundary in the dialect categorization task despite never reaching consistent 75% accuracy in the varying intervals portion of the AX task. The BOUNDARY LOCATION *is* explainable by AX accuracy elsewhere along the continuum.** The upper-left graph is the participant's responses to the dialect categorization task (purple (1) = "Gavaṛ"; yellow (0) = "not Gavaṛ"). The lower-left graph represents the participant's accuracy on AX consistent intervals (dark blue=correct; green=incorrect). The graph on the right represents the participant's responses to AX varying intervals (right column = correct response; left column = incorrect response; colors represent different intervals). The participant's BOUNDARY LOCATION was 2.5.

Figure 6.17 summarizes whether BOUNDARY LOCATION was in an area of relatively high accuracy on the consistent intervals portion of the AX task among those participants who did reach consistent 75% accuracy on the varying intervals portion of the AX task, but whose BOUNDARY LOCATION did not correspond with their SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED from this portion (i.e., among the participants represented by the bright and pale

pink columns in **Figure 6.12**). As noted above, this comparison was carried out only for the words *mas* and *pat*. As the figure shows, for both words there were participants whose BOUNDARY LOCATION was explainable by their AX accuracy (light violet column) and those whose BOUNDARY LOCATION was not explainable by their AX accuracy (dark violet column). **Figure 6.18** provides an example of the responses of an individual participant who represented the dark violet column (i.e., BOUNDARY LOCATION is not explained by AX performance).

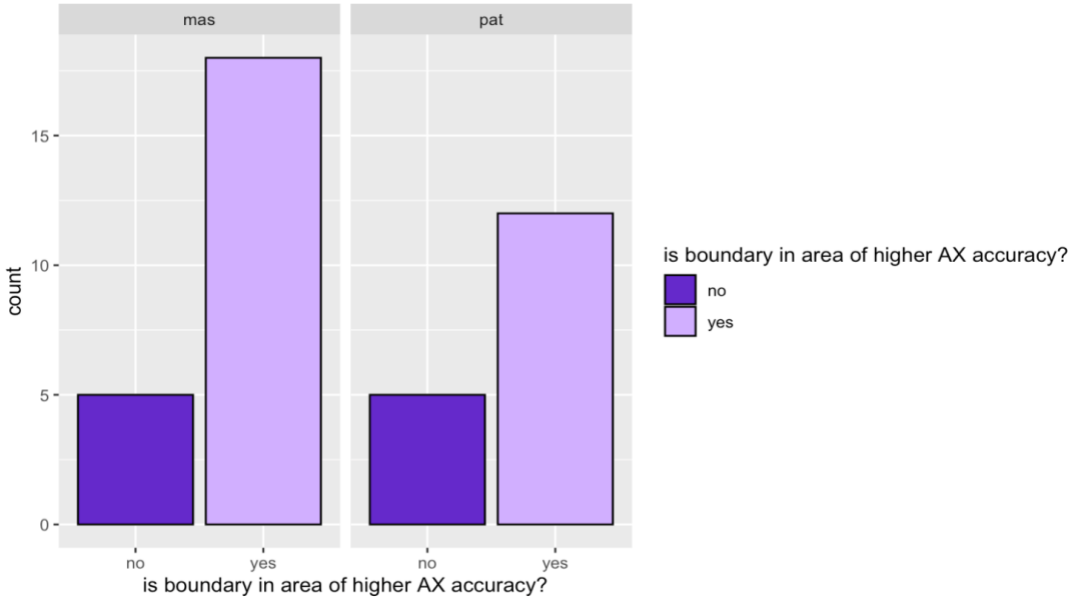


Figure 6.17: Comparison of BOUNDARY LOCATION with areas of relatively high accuracy on the consistent intervals portion of the AX task (among those participants who **reached** consistent 75% accuracy on the varying intervals portion for the relevant word, and whose BOUNDARY LOCATION and SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED did not correspond).

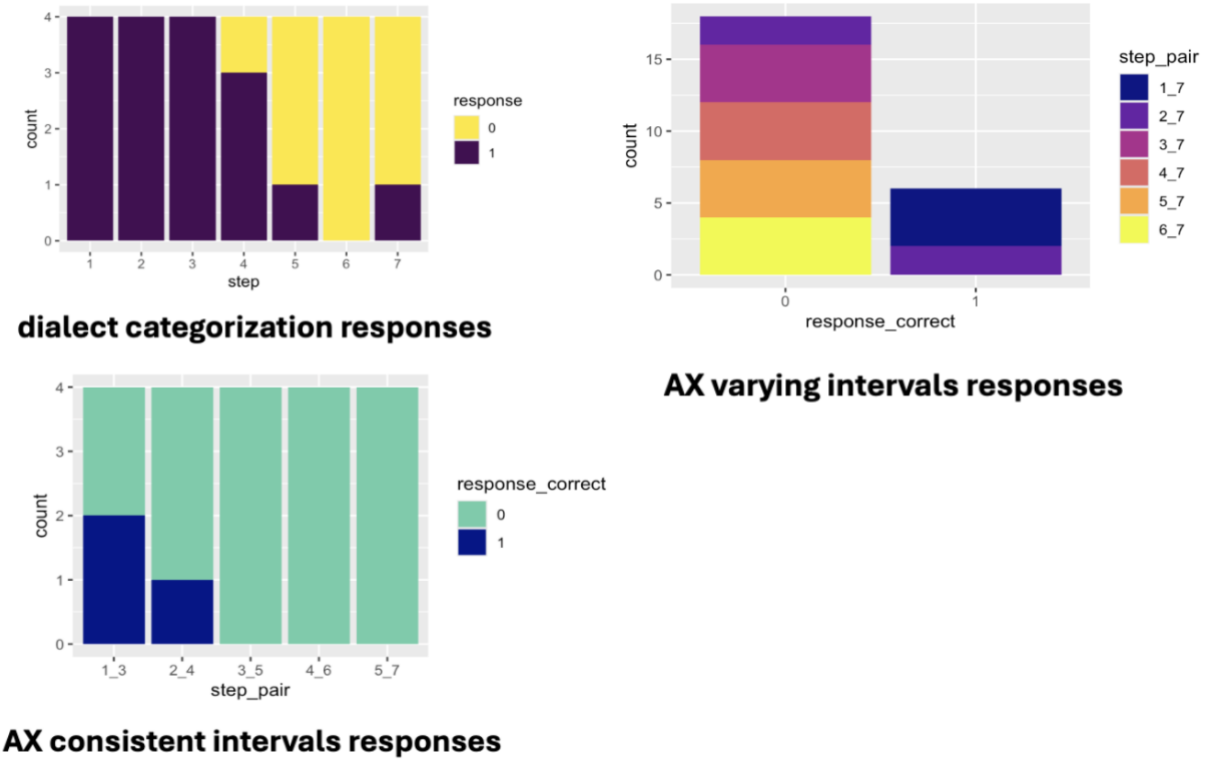


Figure 6.18: Example of participant (Participant 24; *mas*) who **had a perceptual boundary in the dialect categorization task and reached consistent 75% accuracy in the varying intervals portion of the AX task, but whose BOUNDARY LOCATION falls to the right of the SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED from step 7. The SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED was 1-7, while the BOUNDARY LOCATION was 4.76. The boundary location is also not explainable by AX accuracy elsewhere along the continuum**, as the participant had 0% accuracy at the interval 5-7**.** The upper-left graph is the participant's responses to the dialect categorization task (purple (1) = "Gavaṛ"; yellow (0) = "not Gavaṛ"). The lower-left graph represents the participant's accuracy on AX consistent intervals (dark blue=correct; green=incorrect). The graph on the right represents the participant's responses to AX varying intervals (right column = correct response; left column = incorrect response; colors represent different intervals).

Overall, the results indicate that there was a great deal of variation in relationships between performance on the two perception tasks. For example, there were participants who showed good discrimination in the AX task, but nevertheless did not sort the stimuli in a categorical manner in the second task. On the other hand, there were participants who struggled to discriminate stimuli in the AX task, but sorted those same stimuli in a categorical manner in the second task. In general, there were many participants for whom the BOUNDARY LOCATION did not seem related in any way to AX performance (on either the consistent or varying intervals portions of the task), as represented by the dark orange and dark violet groups in the figures above. These results suggest that task type influenced perceptual outcomes (at least for some

participants), such that the social framing of the dialect categorization task activated sensitivity to acoustic differences that were not attended to in the AX task. In this way, the results of this individual-level analysis provide some support for the general result of the comparison of the analyses presented in Chapters 4 and 5 (as discussed in Section 6.3.1 above), according to which the social framing of the dialect categorization task likely facilitated categorization of [a]/[ɒ]$_A$.

Though it is difficult to quantitatively compare the combinations of response patterns across different words due to the small number of participants representing some combinations, there is one trend in the visualizations above that merits further discussion. Namely, the words representing [a]/[ɒ]$_A$ (*pat* and *mas*) appear to show substantially more cases in which participants' BOUNDARY LOCATION corresponded with their SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED from step 7 (see the reddish purple column in **Figure 6.12**), as compared with the words representing [ɔ]/[ɒ]$_O$. There are several potential explanations for this trend. The analysis and visualizations presented in Chapter 5 showed that responses to [a]/[ɒ]$_A$ words in the dialect categorization task generally had steeper SLOPE (and ABSOLUTE VALUE OF SLOPE) values than responses to [ɔ]/[ɒ]$_O$ words, with the word *tsots* in particular having very shallow slopes. Since the boundary locations calculated in relation to very shallow slopes were more likely to be outside of the range of the continuum (i.e., less than 1 or greater than 7), this could possibly explain why less correspondences were found between BOUNDARY LOCATION and SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED for [ɔ]/[ɒ]$_O$. Nevertheless, this explanation may not be sufficient, since the distribution of the steepness of slopes for the word *pos* was relatively similar to those of the words *mas* and *pat* (see **Figure 5.18** in Section 5.6.4.2). An alternative explanation could relate to the phonological status and social salience of [a]/[ɒ]$_A$. As discussed in Chapter 4, the unexpectedly poor AX discrimination of this pair could relate to its allophonic nature. The results of this individual-level analysis (specifically, the relatively large number of cases where participants' BOUNDARY LOCATION for *mas*/*pat* corresponds with their SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED) suggest that for some participants, perception of the vowels of this continuum as dialectal may also have conditioned performance on the AX task. For example, if a participant's SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED was 4-7, and their BOUNDARY LOCATION for the same word in the dialect categorization task was located around step 4, perhaps they could only discriminate the interval 4-7 (and not smaller intervals such as 5-7) because step 4 was the first step that they perceived as a dialect vowel (rather than a

low-level allophonic or coarticulatory difference) in comparison with step 7. In other words, it is possible that perceiving each step as a dialect vs. non-dialect pronunciation could have mediated discrimination in the AX task for some participants, and that discrimination of variants not perceived as dialectal was difficult due to the allophonic/coarticulatory nature of the vowel pair.

Despite the identification of some qualitatively interesting trends in this analysis, it is nevertheless necessary to avoid overinterpreting the results. As discussed in Section 5.6.1 the method of calculating boundary locations (adapted from Butcher 2022) may not be ideal for analysis of the given experimental task, as it was originally designed for analysis of a task that involved phonemic categorization, where the interpretation of shallow or noisy response patterns is different than in the present study. This caveat not only necessitates caution in the interpretation of qualitative trends, but also played a role in the decision to not undertake any quantitative comparison of the combinations of patterns found across different words in this section, as the reliability of such comparisons would be questionable due to the questionable appropriateness of the metric (BOUNDARY LOCATION) upon which they would be based.

6.4.2 Comparison of perception/production with discriminability/merger criteria and individual-level perception-production comparison

6.4.2.1 Questions

The main goal of this individual-level analysis was to determine if participants would likely have been able to discriminate the vowel differences that they themselves and others in the sample produced, and whether or not these differences would be considered discriminable and/or merged by various criteria used in previous studies. The questions investigated were as follows:

1) How do the vowel differences that participants produced compare with criteria for discriminability and/or mergedness used in other studies?
2) How does participants' perception (as judged by AX discrimination) compare with such criteria?
3) Would participants have been likely to discriminate the differences that they themselves produced?

4) Out of the differences produced by all participants for the two relevant vowel pairs, how many of them would have been discriminated by anyone in the sample (i.e., by the best discriminators in the sample)?

It was also explored whether the answers to these questions differed across the two vowel pairs, and whether differences between the two vowel pairs aligned with expectations and with results of other analyses.

6.4.2.2 Methods

6.4.2.2.1 Choosing criteria

As discussed in Section 3.4.1, criteria and methods for determining the size of acoustic difference that is discriminable have sometimes been conflated with criteria and methods for determining the size of acoustic difference that should be considered contrastive vs. "merged." Such criteria and methods are numerous, often arbitrary, and often in disagreement with one another. Among methods that rely on statistical significance testing for defining merger, Stanley and Sneller's (2023) method for creating merger thresholds for Pillai scores improves upon previous methods (discussed in Section 3.4.1), as it considers the effects of differing sample size on Pillai scores. As such, the results of the 2025 production study in comparison with these thresholds will be briefly restated in the results section below.

The other category of methods mentioned in Section 3.4.1 consists of attempts to come up with acoustic thresholds for vowel formants (in Bark or Hz). In some studies, such thresholds are explicitly intended to evaluate discriminability of vowels (i.e., without making claims about phonological states such as "merger"), while other authors attempt to establish thresholds for merger and/or near-merger, sometimes nevertheless relying on previous studies that investigated only discriminability. For the purposes of the present analysis, such methods have an advantage over significance testing because they can be used to compare participants' perception and production. One author who explored the question of acoustic discriminability of vowels without approaching the question of phonological contrastiveness is Kewley-Port (2001). Based on the results of a series of experiments involving small manipulations of vowel formant frequencies, Kewley-Port (2001: 2153) established 0.5 Bark "as a typical threshold for formant discrimination

by listeners with little training in a variety of experimental conditions," noting also that the discrimination threshold may be closer to 0.3 Bark in "normal discourse" (2001: 2154).

In other studies, the issue (discussed in Section 3.4.1) of conflating acoustic discriminability with phonological contrastiveness is apparent. For example, Kotsoni (2019) defines thresholds for merger, near-merger, and contrast on the basis of thresholds used by Recasens and Espinosa (2009), which in turn come from "just noticeable difference" data reported by Boersma (1998: 104), which originally comes from Flanagan (1955). The goal of Flanagan's study was similar to that of Kewley-Port's – in other words, the study involved experiments with the goal of determining the size of differences in vowel formant frequencies that are detectable to listeners. Crucially, Flanagan and Kewley-Port used different criteria for defining the success rate that equated to a difference being discriminable – Kewley Port's discrimination thresholds were defined as 71% correct responses (2001: 2144), while Flanagan's were defined at 50% (1955: 616). This could explain why the formant thresholds reported by Flanagan are smaller (around 0.2 Bark, when converted from the Hz values reported by Flanagan) than those reported by Kewley-Port. However, the list of subsequent authors who base their merger / near-merger / contrastiveness thresholds on the numbers reported by Flanagan do not discuss the implications of Flanagan's methods and/or the suitability of borrowing such thresholds for studies that are concerned with phonological aspects of perception. In general, there seems to be no a priori reason for assuming that discriminability thresholds calculated with either method (71% correctness vs. 50%) would have any direct relationship to phonological contrastiveness/mergedness, or even to discriminability of acoustic differences in cases where certain phonological factors (e.g., language-specific allophony) are at play. Thus, in comparing such thresholds to one's perception and production data, it is necessary to be clear about what the given studies actually intended to investigate, in comparison with the goals of one's own study.

A different approach for acoustically defining merger in production was taken by Evanini (2009: 141-142). As Evanini's study involved data sets where both production and perception data was available for some individuals, but only production data was available for others, he sought to identify an acoustic threshold for merger in production that could be used to evaluate the mergedness of those speakers for whom only production data was available. In discussing production data from the Atlas of North American English, Evanini notes that the Euclidean distance between the vowel categories involved in the English low-back vowel merger is less

than 100 Hz in regions that are known to have this merger, thus taking 100 Hz as a production threshold below which individuals for whom only production data was available could be deemed to also have the merger. In other words, Evanini's strategy involved observing what kind of acoustic difference between the low back vowels was present in the speech of those who had already been deemed to be phonologically merged based on a combination of analyses (including perceptual analyses such as minimal pair tests), and then applying the derived metric (100 Hz) to identify individuals with no perception data as either merged or unmerged in production. While this method is somewhat circular, it at least has the advantage of using data related to perception of phonological contrastiveness (e.g., data from minimal pair tests) in order to define an acoustic threshold for phonological contrastiveness in production, rather than using questionably applicable data related to acoustic discriminability.

In order to obtain a thorough picture of how participants' produced and discriminated differences in the present study compare to various types of criteria, the results of several analyses are presented below. As mentioned above, the size of participants' produced vowel differences in comparison with Stanley and Sneller's (2023) thresholds will be briefly discussed. As Kewley-Port's (2001) discriminability threshold of 0.5 Bark in a given formant (for untrained listeners in experimental contexts, based on 71% correctness) seems suitable given the conditions of the present study (experiment with untrained listeners, for which a threshold of 75% accuracy was used to define one metric of AX performance), participants' AX discrimination and production are compared with this threshold. Participants' AX discrimination and production are also compared with the threshold of 0.2 Bark for a given formant, which Recasens and Espinosa (2009) (based on the discriminability thresholds reported by the aforementioned previous authors) define as unmerged (while differences smaller than this are defined as near-merged). While comparisons with the threshold of 0.5 and 0.2 Bark are carried out on a per-formant basis, participant's AX discrimination and production as quantified by Euclidean distances between their perceived and produced categories are compared with the 100 Hz threshold for merger in production reported by Evanini. In order to convert this threshold to Bark (i.e., in order to determine approximately how many Bark is equal to 100 Hz in the area of the vowel space explored by Evanini), the following method was used. First, average formant values for the two US English vowel categories examined by Evanini (/ɑ/ and /ɔ/) for men and women were obtained from Peterson and Barney (1952). According to the data reported by

Peterson and Barney (1952), the Euclidean distance between the average formant values of the two categories was about 300 Hz for men and 400 Hz for women (it is not relevant whether Peterson and Barney's participants, individually or in general, would be deemed "merged" or "unmerged" with respect to this vowel contrast, as the goal of using these formant values was simply to determine approximately how many Bark 100 Hz is equal to in the low back region of the vowel space, rather than to comment on US English speakers' maintenance of the low back contrast). For each gender group, the distance between the endpoint categories (i.e., between Petersen and Barney's reported formant values for the two categories) was divided into 100 Hz segments (3 segments for men and 4 segments for women). Each 100 Hz segment was converted to Bark, and the resulting values were averaged, with a resulting Euclidean distance of around 0.65 Bark. Thus, Euclidean distance between Gavaṛ participants' perceived and produced categories is compared with this threshold, in order to assess how the differences perceived and produced in the present study compare with Evanini's threshold for merger in production.

6.4.2.2.2 Comparing thresholds, perception, and production via visualizations

Comparison of participants' discriminated and produced differences with the thresholds of 0.5 Bark (per formant), 0.2 Bark (per formant), and 0.65 Bark (Euclidean distance) were carried out using data visualizations. Raw Bark values (representing both participants' produced differences and the differences presented in the perception stimuli) were used, as the goal of this analysis was to investigate acoustic discriminability, rather than, for example, to compare the locations of different participants' vowel categories in the vowel space (which would require normalized data). Participants' AX discrimination was represented by the SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED metric described in Chapter 4. Visualizations were also used to compare participants' discrimination with their own productions, and to compare participants' productions with the perception of the best discriminators in the sample. Participant 16 was excluded from all analyses involving perception data (as this participant's perception data was lost – see Section 3.5.2), while Participant 36 was excluded from analyses involving the results of the AX task (see Section 4.4.1). An example of the visualizations used in this analysis is presented in **Figure 6.19**. Participants' produced and perceived Bark values were rounded to 2 decimal places, and in comparing them with the various thresholds, the values were divided into those that were less than a given threshold vs. equal to or greater than the threshold. For

example, if a participant produced a difference of exactly 0.5 Bark, then in comparison with the discriminability threshold of 0.5 Bark, this difference would be considered discriminable.

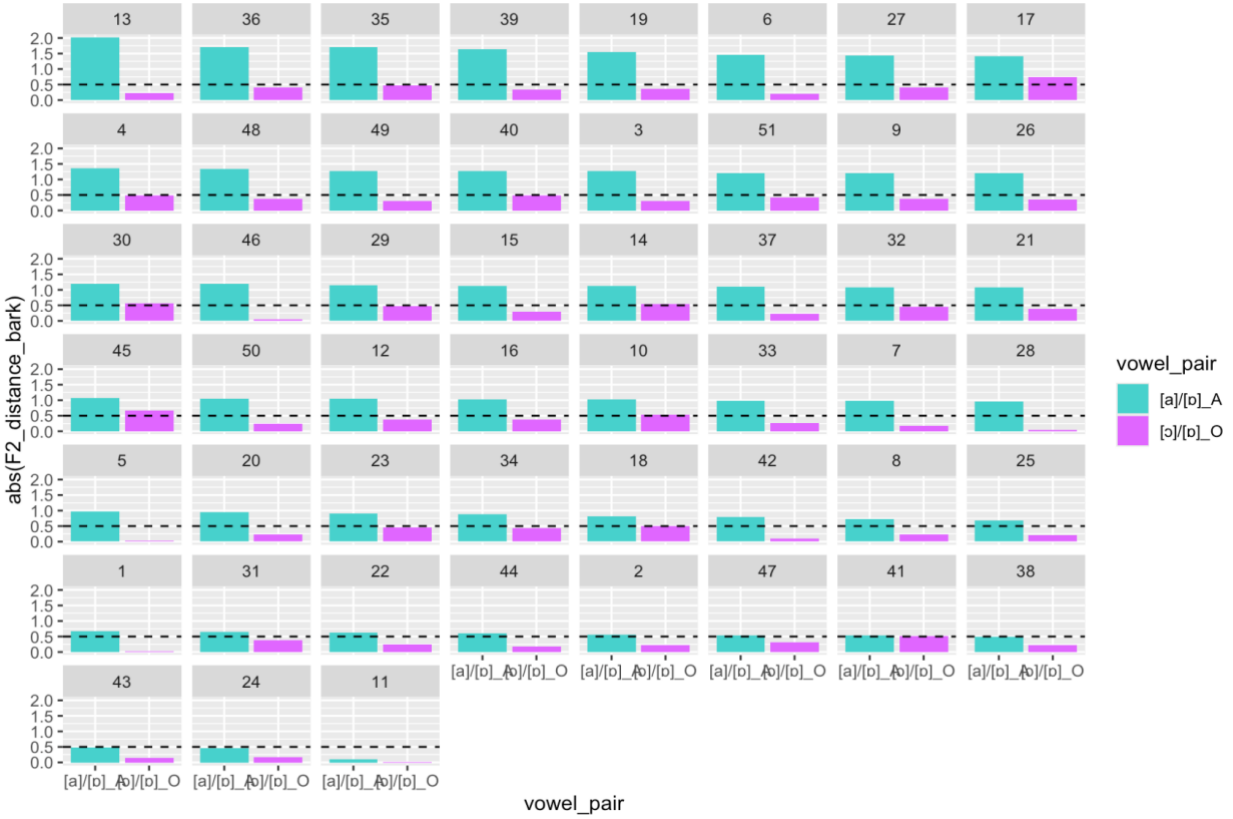


Figure 6.19: Comparison of individual participants' produced differences in F2 (in raw Bark values) with Kewley-Port's (2001) discriminability threshold of 0.5 Bark. The threshold is represented by the horizontal dotted line, while the two columns represent the difference in F2 between the members of each vowel pair in each participant's production data. Participants are ordered by their value of F2 for [a]/[ɒ]$_A$, from highest to lowest.

6.4.2.3 Results

Figures 6.20-6.21 below present participants' 2025 Pillai scores for [a]/[ɒ]$_A$ and [ɔ]/[ɒ]$_O$ with overlaid thresholds calculated by the method of Stanley and Sneller (2023). As the figures show, all participants have a score deemed distinct by these thresholds for [a]/[ɒ]$_A$, and around two-thirds of participants have scores deemed distinct for [ɔ]/[ɒ]$_O$. While this difference between the two vowel pairs might initially seem surprising given the contrastive nature of [ɔ]/[ɒ]$_O$ (as reported by dialectological sources such as Katvalyan [2016]), the results and discussion presented in Chapters 2 and 3 touch on several possible explanations for the maintenance of larger and/or more consistent differences in [a]/[ɒ]$_A$. As discussed in Chapter 2, the gender-based

patterns of variation found for this vowel pair in production were more consistent across the 2022 and 2025 data sets in comparison with [ɔ]/[ɒ]$_{O}$, indicating that it could be maintained more consistently as a dialect feature and used more consistently to express social meaning (which would also be unsurprising given its more prominent social salience). Furthermore, as discussed in Chapter 3, the data from Yerevan Armenian presented by Seyfarth and Garellek (2018) suggest that speakers in Yerevan exhibit small (possibly coarticulatory) differences from [a] in phonetic contexts that are similar to those conditioning the [a]/[ɒ]$_{A}$ allophony in Gavaṛ. Given the increasing influence of Yerevan Armenian throughout the country, it is thus possible that speakers who do not exhibit the very large allophonic differences characteristic of Gavaṛ's dialect have converged towards Yerevan's system (i.e., such speakers should not be expected to show complete overlap of [a] and [ɒ]$_{A}$, because small differences between these vowels can be found in the influential supralocal variety that they are likely converging toward).

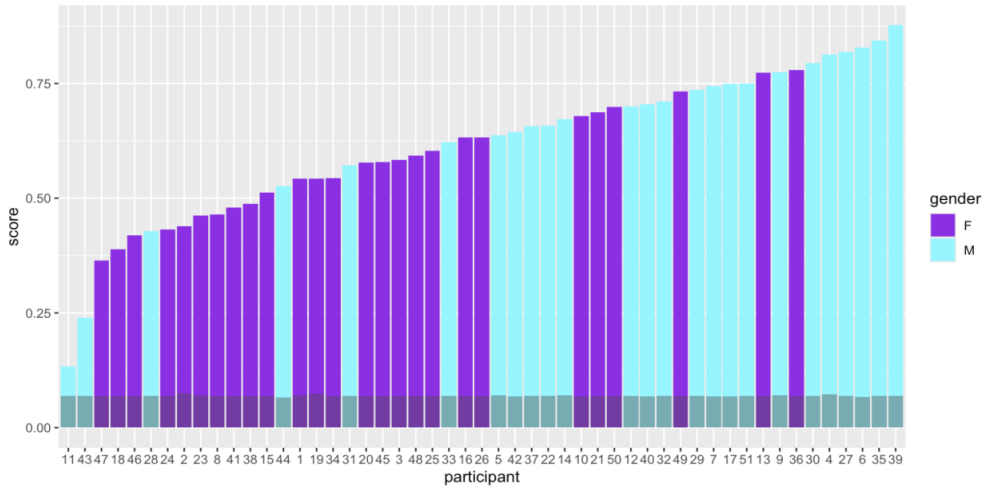


Figure 6.20: Individual Pillai scores by GENDER for **[a]/[ɒ]$_{A}$** with superimposed distinction thresholds. Participants are ordered from lowest Pillai score (most overlapped vowels) to highest Pillai score (least overlapped vowels).

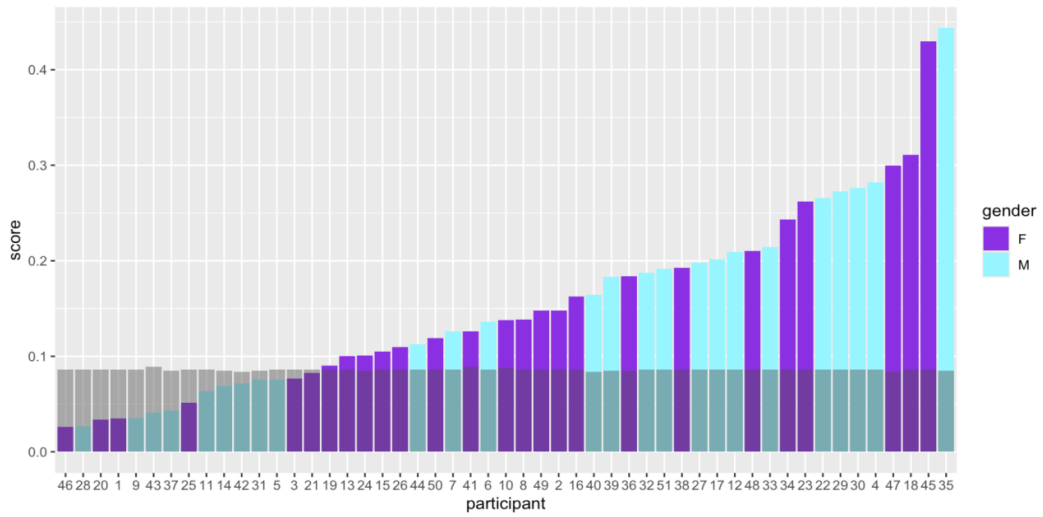


Figure 6.21: Individual Pillai scores by GENDER for **[ɔ]/[ɒ]$_O$** with superimposed distinction thresholds. Participants are ordered from lowest Pillai score (most overlapped vowels) to highest Pillai score (least overlapped vowels).

Table 6.1 below presents the number of participants whose productions of the two relevant vowel pairs are deemed discriminable and/or unmerged by the three threshold methods described in Section 6.4.2.2. The table also lists the number of participants whose productions were indiscriminable/merged according to the given threshold method.

Table 6.1: Comparison of participants' produced vowel differences for [a]/[ɒ]$_A$ and [ɔ]/[ɒ]$_O$ with the three Bark thresholds described in Section 6.4.2.2. The first two rows show the number of participants whose produced F1 and F2 differences were larger than the 0.5 and 0.2 Bark per-formant thresholds (i.e., were discriminable/unmerged). The third row shows the number of participants for whom the Euclidean distance between the relevant vowels in production was larger than the 0.65 Bark threshold (i.e., were unmerged). The fourth row shows the number of participants whose productions would have been deemed indiscriminable/merged by each method.

	0.5 Bark		**0.2 Bark**		**0.65 Bark**	
	[a]/[ɒ]$_A$	**[ɔ]/[ɒ]$_O$**	**[a]/[ɒ]$_A$**	**[ɔ]/[ɒ]$_O$**	**[a]/[ɒ]$_A$**	**[ɔ]/[ɒ]$_O$**
F1 produced difference larger	10	0	26	15	n/a	n/a
F2 produced difference larger	47	7	50	41	n/a	n/a
Euclidean distance larger	n/a	n/a	n/a	n/a	43	2
# participants not discriminable / not unmerged in either formant	4	44	1	9	8	49

Table 6.2 compares the same three Bark thresholds to participants' discrimination of the 4 words used as stimuli in the AX task. For the purposes of this analysis, the smallest difference that each participant perceived was represented by the SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED (at which they achieved consistent 75% accuracy for the relevant word in the varying intervals portion of the AX task). As the table shows, participants' perceived F2 differences for *pos* and *tsots* were not compared with the thresholds, as it was determined that this comparison would be uninformative due to the fact that the F2 differences presented in these stimuli were generally very small in comparison with other presented differences (see **Table 4.6** in Section 4.3.2.1). As such, with respect to this formant for the words *pos* and *tsots*, participants essentially had no choice but to perceive very small differences (for example, for both words, the F2 difference in all participants' SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED was less than 0.5 Bark). As the F1 differences presented in the stimuli for this vowel pair were larger and more varied (i.e., they could be either large or small), examining participants' discrimination of this formant is more informative (while for [a]/[ɒ]$_A$, the stimuli had substantial differences in both formants, making them both informative).

Table 6.2: Comparison of participants' SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED for each word with the three Bark thresholds described in Section 6.4.2.2. The first two rows show the number of participants whose SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED was smaller in F1 or F2 than the 0.5 and 0.2 Bark per-formant thresholds (F2 is not considered for *pos*/*tsots*, for the reasons described above). The third row shows the number of participants for whom the SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED (Euclidean distance) was smaller than the 0.65 Bark threshold.

	0.5 Bark				**0.2 Bark**				**0.65 Bark**			
	mas	***pat***	***pos***	***tsots***	***mas***	***pat***	***pos***	***tsots***	***mas***	***pat***	***pos***	***tsots***
F1 perceived difference smaller	0	0	14	6	0	0	0	0	n/a	n/a	n/a	n/a
F2 perceived difference smaller	0	0	n/a	n/a	0	0	n/a	n/a	n/a	n/a	n/a	n/a
Euclidean distance smaller	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	0	0	30	6

Finally, **Table 6.3** presents the number of participants who would have discriminated the differences in their own productions (as quantified by raw Bark Euclidean distance between the

relevant vowel categories). For the purposes of this analysis, the smallest differences that each participant perceived were represented by their SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED from the AX task. For each vowel pair, a participant was considered to have likely been able to discriminate their own produced difference if their produced difference was larger than their SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED (Euclidean distance) for either of the words representing the given vowel pair. This table also presents the results of the analogous analysis with respect to the perception of the best discriminators in the sample

Table 6.3: The first row shows the number of participants who likely would have discriminated their own produced differences for each vowel pair, if the smallest perceived difference is defined as the SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED in the varying intervals portion of the AX task. The second row shows the number of participants whose produced differences for each vowel pair likely would have been discriminated by anyone in the sample, according to the same analysis method.

	[a]/[ɒ]$_A$	**[ɔ]/[ɒ]$_O$**
Discriminable by self	5	6
Discriminable by anyone in sample	15	29

6.4.2.4 Discussion and limitations

As the figures and tables above show, the answer to the question of how many participants produced discriminable and/or "unmerged" differences between the relevant vowels differs vastly according to different methods. According to some methods (Stanley & Sneller [2023] thresholds; 0.2 Bark threshold), all or most participants had distinct/unmerged vowels for both vowel pairs, while according to other methods, most participants had a distinction in [a]/[ɒ]$_A$, but not in [ɔ]/[ɒ]$_O$. According to all methods, the number of participants with discriminable/unmerged vowels was greater for [a]/[ɒ]$_A$ than for [ɔ]/[ɒ]$_O$, which is unsurprising given that the acoustic differences (in raw Bark values) produced by participants were generally larger for [a]/[ɒ]$_A$ (this is also reflected by the Pillai scores shown in **Figures 6.20-6.21**; see Section 6.4.2.3 above for discussion of potential explanations for the maintenance of larger differences in [a]/[ɒ]$_A$).

The results of the analysis presented in the previous section align with the results presented in Chapter 4 in suggesting that discrimination of [a]/[ɒ]$_A$ was unexpectedly poor, considering that the differences between the members of this vowel pair were larger than for [ɔ]/[ɒ]$_O$ both in the stimuli and in the participants' own productions. As **Table 6.2** shows, for the words *mas* and *pat*, no participants' SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED from the varying intervals portion of the AX task was smaller than any of the three thresholds tested. On the other hand, for *pos* and *tsots*, there were participants who discriminated differences smaller than two out of the three thresholds. The SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED of any participant for [a]/[ɒ]$_A$ was 1.27 Bark (Euclidean distance between the given continuum steps), while for [ɔ]/[ɒ]$_O$ the smallest interval was 0.35 Bark. Given how well some participants performed at discriminating *pos*/*tsots*, it might have been expected that some participants would also discriminate the 1-step interval (6-7) for [a]/[ɒ]$_A$, which was slightly larger than the smallest discriminated differences for [ɔ]/[ɒ]$_O$ (around 0.43 Bark for the word *pat*). However, no participant consistently discriminated the interval 5-7 for [a]/[ɒ]$_A$, let alone 6-7. Thus, the results presented in **Table 6.2** support the interpretation that phonological factors (e.g., allophony) inhibited discrimination of [a]/[ɒ]$_A$, whose members were much more acoustically distinct than [ɔ]/[ɒ]$_O$ in both the talker's and participants' pronunciations. The analysis presented in **Table 6.3** also supports this interpretation, as it suggests that the number of participants whose produced differences would have been discriminated by themselves or by anyone in the sample is either similar across the two vowel pairs, or is greater for [ɔ]/[ɒ]$_O$, though a greater number would have been expected for [a]/[ɒ]$_A$ based on the size of acoustic differences alone. In combination with the comparison of participants' productions with the various thresholds, the results presented in **Tables 6.2-6.3** suggest that produced differences in [a]/[ɒ]$_A$ are potentially more discriminable than those in [ɔ]/[ɒ]$_O$, but are not actually discriminated more successfully.

Nevertheless, the interpretation of the above results is limited by the fact that the varying intervals portion of the AX task did not account for discrimination of varying-size intervals at different points along the vowel continua. In other words, the above analysis assumes that participants' discrimination of differences from the canonical SEA variant is somehow more relevant for comparison with various thresholds and with participants' productions than discrimination of differences of similar size elsewhere along the continua. This assumption is not unfounded – for example, all speakers produce the pronunciation [a] in at least some phonetic

contexts, while not all speakers produce strongly dialect-like [ɒ]$_A$ pronunciations, so at least in the case of comparison with participants' productions, one could argue that discrimination of differences from [a] inform us about a more relevant area of the vowel space in comparison with other parts of the continuum. On the other hand, it cannot necessarily be said that participants' discrimination of differences from the talker's [a] pronunciation tell us about how they would perceive differences from their own or other participants' [a] pronunciations, as the exact location of each speaker's [a] (in terms of raw Hz values of formants, and in terms of location relative to the speaker's other vowels) will be slightly different.

As discussed in Chapter 4, in the case of the [a]/[ɒ]$_A$ continuum in particular, many participants were more successful at discriminating 2-step intervals elsewhere along the continuum than they were at discriminating the interval 5-7 (which represents a small difference from the pronunciation [a], and which was not discriminated by any participant for *pat* or *mas* with 75% consistent accuracy). Thus, if participants' discrimination of differences at different locations along the continuum had been considered in the analysis above, the results would have been very different. For example, if we consider participants who achieved 75% accuracy at some other 2-step interval (i.e., not 5-7) for *pat* or *mas*, then the number of participants who would have discriminated their own produced difference would be 31, rather than 5. Unfortunately, it is not possible to adjust the analysis in an analogous way for [ɔ]/[ɒ]$_O$, as the structure of the task was such that no data is available for 2-step intervals other than 5-7. Likewise, if the other 2-step intervals for [a]/[ɒ]$_A$ are taken into consideration, then 36 participants' produced differences in [a]/[ɒ]$_A$ would have been perceived by the best discriminators in the sample (rather than 15). However, again, we cannot analyze the [ɔ]/[ɒ]$_O$ perception data in the same way due to the structure of the task. Thus, while the results of the above analysis are limited by the fact that they tell us about discrimination of varying size intervals only in a specific area of the vowel continua, it is not possible to carry out the same analysis in relation to discrimination of intervals at different points along the continua.

6.5 General summary and discussion

6.5.1 Comparison of performance on the two perception tasks

The most notable finding related to comparison of perception across the two tasks comes from the general comparison of the results for the two vowel pairs. Specifically, broad

comparison of the findings shows that discrimination was unexpectedly poor for [a]/[ɒ]$_{A}$ given the sizes of the differences presented in the stimuli, but that participants nevertheless categorized the stimuli for this vowel pair in a more expected manner than the stimuli for [ɔ]/[ɒ]$_{O}$ in the second perception task. Thus, while the results defy what one would expect if the main factor affecting perception were the size of acoustic differences presented in stimuli, they align with other expectations formed on the basis of the characteristics of the two vowel pairs. In particular, the allophonic nature of [a]/[ɒ]$_{A}$ (and perhaps also the coarticulatory nature of small differences from the SEA variant [a]) seem to have impeded discrimination of this vowel pair in the AX task, while the social salience of this vowel pair may have facilitated its categorization in the second task, where the social framing of the task may have allowed participants to attend to differences that they did not discriminate in the AX task.

The results of the direct quantitative comparison of performance on the two perception tasks presented in Chapter 5 are more difficult to interpret. In the statistical models that involved predicting individual-level metrics of performance on the dialect categorization task with individual-level metrics of performance on the AX task, some weak evidence was found suggesting that accuracy in the AX task might relate to categorizing stimuli more categorically and/or in a more expected manner in the dialect categorization task. However, as noted above, these effects were limited, weak (i.e., not always significant), and sometimes contradicted other weak trends found in the data. The analyses presented in Chapter 5 thus do not tell us anything conclusive about relationships between individual performance across the two perception tasks.

The more qualitative analysis of individual-level performance across the two perception tasks presented in this chapter reveals a vast amount of individual variation in patterns of performance across the two perception tasks. For example, some participants had relatively accurate discrimination in the AX task, but did not sort the stimuli categorically and/or in the expected manner in the second task (see **Figure 6.11**), while other participants had poor discrimination in the AX task but sorted the stimuli categorically and/or expectedly in the second task (see **Figure 6.12**). There was also vast variation in terms of the relationship between BOUNDARY LOCATION in the dialect categorization task with various aspects of performance in the AX task. For some participants, the location of the perceptual boundary seemed related to AX performance (on either the consistent or varying intervals portions of the task), while for others, no relationship could be discerned (see **Figures 6.12, 6.14, 6.17**). Thus, at the individual

level, it seems that task type can affect perceptual outcomes differently for different people, and that performance across the two perception tasks was transparently related for some participants but not for others. While some participants' individual results point to the same tendency observed in the overall comparison of the results for the two vowel pairs (i.e., the social framing of the dialect categorization task activating sensitivity to small acoustic differences that were not necessarily discriminated in the AX task), there are also individual participants whose performance across the two perception tasks (at least for some stimuli words) did not align with this tendency (see, e.g., **Figure 6.11**). The degree of individual variation found (including the opposite tendencies exhibited by some participants in terms of which task they performed more accurately on) could perhaps be related to the general complexity of the linguistic situation in Gavaṛ. As shown by the 2022 and 2025 production results, Gavaṛ residents vary widely in terms of their maintenance of differences in the two dialectal vowel pairs explored in the perception experiment. If some degree of dialect levelling is currently occurring (as proposed by Hodgson [2019] and Weitenberg [2002] – see Chapter 1), such that Gavaṛ residents are regularly exposed to a variety of different vowel systems, it is reasonable to assume that their exposure to the maintenance of dialect-like differences between the vowels and their metalinguistic awareness of dialect features would also vary. Thus, the individual variation found in participants' responses to the two perception tasks could be related in part to their heterogeneous and fluid linguistic environment.

As noted above, the only clear trend observed in the comparison of the two vowel pairs in the given analyses was that the words *mas*/*pat* had a relatively large number of participants whose BOUNDARY LOCATION in the dialect categorization task corresponded with their SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED from the AX task. This could suggest that for some participants for the vowel pair [a]/[ɒ]$_A$, the variants perceived as discriminable from the SEA variant [a] align with those perceived as dialectal. Such a tendency could be explained by the allophonic and socially salient nature of [a]/[ɒ]$_A$, such that some participants might have mainly been able to discriminate differences that corresponded with their social categories of dialectal vs. non-dialectal variants. Nevertheless, as discussed above, it is important not to overinterpret this result, especially given that the method used for the calculation of perceptual boundaries (based on Butcher [2022]) may not have been ideally suited for analysis of the given data.

6.5.2 Comparison of perception and production, and implications for defining merger

The general comparison of demographic factors relevant for production vs. perception found that different factors were predictive of variation in each analysis (production vs. AX perception vs. perception in the dialect categorization task). While GENDER was the primary factor predictive of variation in production, the analysis of the AX task (Chapter 4) found variation in relation to BIRTH YEAR and EDUCATION LEVEL, while the analysis of the dialect categorization task (Chapter 5) found variation in relation to BIRTH YEAR and GENDER. Notably, while men were found to produce more dialect-like variants of some vowels in production, they were found to categorize stimuli less accurately in the dialect categorization task. Thus, the findings suggest that the social factors conditioning production and perception are not necessarily the same, and also suggest that the social groups using certain features in production are not necessarily the most attuned to those features in perception. Sociolinguistic studies have generally found that women lead sound changes, adopting prestigious forms before men in "change from above" (the level of consciousness) and adopting innovative variants before men in "change from below" (the level of consciousness) (Labov 2001: 272-293). Thus, in the case of the present study, perhaps women's less robust use of some dialect vowels in production indicates adoption of a change from above (in this case, possible dialect leveling), while their more accurate categorization of dialect vowels in perception could be indicative of increased metalinguistic awareness, which could in turn play a role in their generally faster adoption of sound changes (especially "from above") in production.

As noted above, findings from the quantitative analysis of production metrics as predictors of performance on the two perception tasks were very limited. The most notable finding was the sub-significant tendency for participants with more dialect-like productions (greater distances, less overlap between the relevant vowels) to have perception boundaries further toward the dialect-like end of the continuum for [a]/[ɒ]$_{A}$, which could suggest a link between the size of produced differences and the size of the difference needed for one to perceive the dialect variant. This effect was nevertheless very weak, such that it is difficult to draw strong conclusions based on it.

The general lack of findings related to production-perception links in the quantitative analysis could potentially be explained by the need for listeners to understand the speech of both those who maintain the dialect vowel system and those who do not (regardless of what their own productions are like). As discussed in Chapter 3, such an explanation has been invoked by many

previous authors in cases where listeners are perceptually attuned to features that they themselves do not use in production but which are widespread in the speech community. Beddor (2015) proposes that the opposite type of individual ("reliable producers / insensitive perceivers") is rare. However, examination of the results of some participants of the present study suggests that such individuals exist and likewise contributed to the lack of quantitative correlations found between production and perception. **Figures 6.22** and **6.23** below show the individual data for [a]/[ɒ]$_A$ from both perception tasks for Participant 39 and Participant 35, who had the first and second largest Pillai scores (i.e., least overlapped vowels, equating to the most dialect-like pronunciations) for this vowel pair. Participant 39 discriminated this vowel pair with very high accuracy in the AX task, but never categorized it in the expected manner in the dialect categorization task. On the other hand, Participant 35 had more difficulty discriminating this vowel pair (for example, discriminating almost no intervals whatsoever for the word *pat*), but nevertheless categorized all three of the words representing it in the expected manner in the dialect categorization task. Thus, both of these participants (who produced some of the largest differences between the given vowels) struggled to recognize and use such differences in perception. While these results illustrate the existence of "reliable producers / insensitive perceivers," they also suggest that use of a feature in production might correlate with perception of that feature in one type of task but not in another type of task, and that such relationships can vary across individuals.

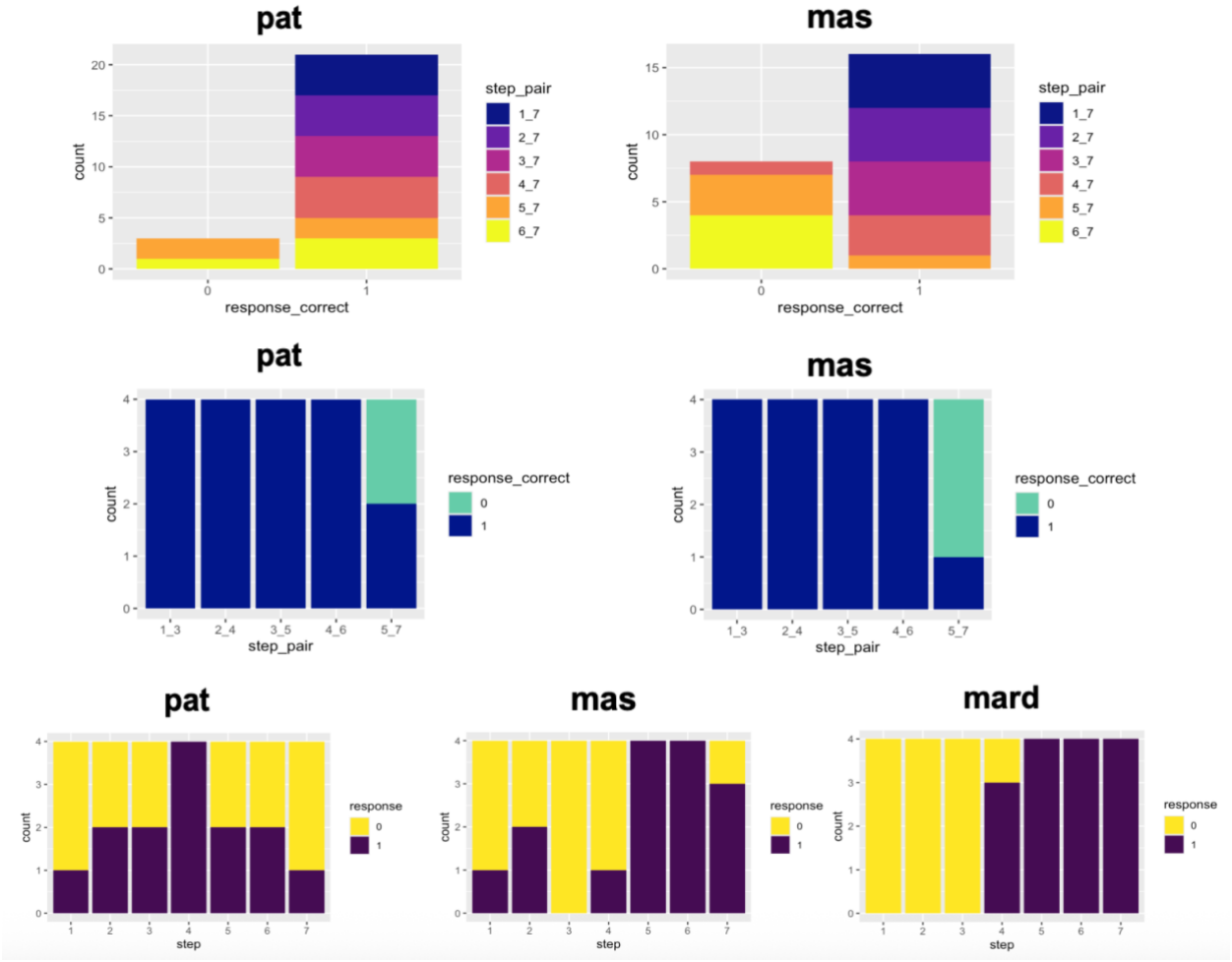


Figure 6.22: Individual perception results for [a]/[ɒ]$_A$ of Participant 39, who had the largest Pillai score for this vowel pair. The top row represents AX varying intervals (left column = incorrect response; right column = correct response, i.e., successful discrimination). The middle row represents AX consistent intervals (dark blue (1) = correct response, i.e., successful discrimination; green (0) = incorrect response). The bottom row represents the dialect categorization task (purple (1) = "Gavaṛ"; yellow (0) = "not Gavaṛ").

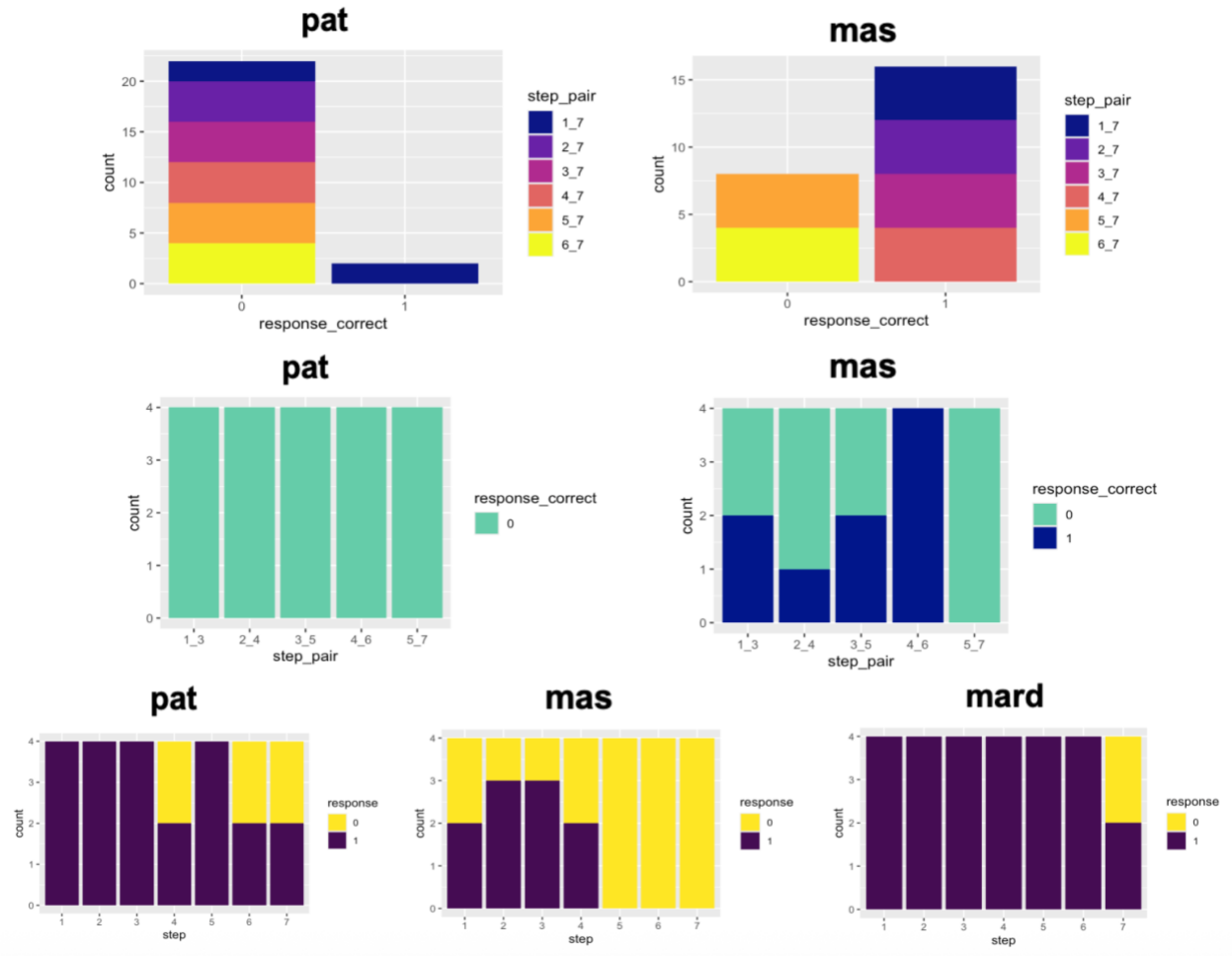


Figure 6.23: Individual perception results for [a]/[ɒ]$_A$ of Participant 35, who had the second largest Pillai score for this vowel pair. The top row represents AX varying intervals (left column = incorrect response; right column = correct response, i.e., successful discrimination). The middle row represents AX consistent intervals (dark blue (1) = correct response, i.e., successful discrimination; green (0) = incorrect response). The bottom row represents the dialect categorization task (purple (1) = "Gavaṛ"; yellow (0) = "not Gavaṛ").

The qualitative perception-production analyses presented in this chapter, though limited in scope because of the limited types of intervals presented in the AX task, provide many insights into relationships between production and perception at the individual and community levels, as well as insights into the implications of the present study for discussions of what it means for vowels to be distinct vs. merged. According to all explored methods for determining discriminability and/or mergedness, the number of participants who produced discriminable/unmerged vowels was greater for [a]/[ɒ]$_A$ than for [ɔ]/[ɒ]$_O$, with the difference between the two vowel pairs being substantial for some of the methods. Despite participants generally producing differences in [a]/[ɒ]$_A$ that should in theory be discriminable based on

numerous previous methods of determining this, the results of the AX task suggest that the general threshold for discriminating these vowels (but not necessarily for discriminating the contrastive vowel pair [ɔ]/[ɒ]$_O$) in the community is much larger than the thresholds proposed in previous literature (even a 2-step interval for [a]/[ɒ]$_A$ was around 0.85 Bark, and as the discussion above notes, many participants were unsuccessful at discriminating such intervals either in general or at certain areas of the continuum). As such, even for the [a]/[ɒ]$_A$ pair (for which most participants produced relatively large differences), a large portion of participants would have been unlikely to discriminate their own produced differences, and many of the theoretically discriminable differences produced by participants would likely not have been perceived by anyone in the sample. The results of these analyses thus support the hypothesis that some participants would produce measurable (i.e., by various criteria unmerged and/or discriminable) differences that are smaller than the smallest differences that they themselves or any of the other participants would perceive as different (at least in certain tasks / circumstances). The fact that such a tendency is more pronounced in the case of [a]/[ɒ]$_A$ aligns with the phonological characteristics of this vowel pair, as the allophonic and/or coarticulatory nature of these vowel differences could be inhibiting discrimination of small differences, especially in certain areas of the vowel continuum.

The overall results of the present study suggest that the size of vowel differences that are discriminable is highly context dependent (i.e., depends not only on phonological characteristics of vowels, but also on their social characteristics and on the circumstances [e.g., task type] under which they are perceived). This in turn suggests that characteristics of specific vowel phenomena, along with local perception of these phenomena, should be incorporated into context-dependent definitions of distinctness and merger. In cases like the [a]/[ɒ]$_A$ phenomenon explored in this dissertation, where vowel differences are conditioned by phonetic context (with potential involvement of coarticulation), the results suggest that listeners (at least under some circumstances, such as the AX task) do not perceive the smallest differences between the given vowels, even if these are differences that previous research would deem to be likely discriminable. Though the social framing of the dialect categorization task appears to have facilitated perception of differences in [a]/[ɒ]$_A$ for some participants, the results of both the AX task and dialect categorization task (in particular, the average location of perceptual boundaries for [a]/[ɒ]$_A$ being close to the middle of the continuum, and further left than the boundaries for

[ɔ]/[ɒ]$_{O}$) suggested that perception of small differences is particularly inhibited near the SEA-like variant [a].

Because of the allophonic and potentially coarticulatory nature of the phenomenon, the fact that speakers maintain acoustic differences between these vowels in production without necessarily being able to perceive such differences is not as mysterious as some previously identified cases of near-merger of contrastive vowels. In other words, in contrast with some previously identified cases of near-merger, the fact that the [a]/[ɒ]$_{A}$ vowels might never reach full merger (as defined by lack of consistent acoustic differences) is not surprising given the phonological characteristics of the vowels involved. Nevertheless, Evanini's (2009) discussion of the English low-back vowel merger (see Section 3.4.1.1) raised the important point that the phonetic contexts surrounding vowels of different classes are not always evenly distributed in the lexicon even in cases of contrastive vowels. Though it is possible to control for phonetic context in analyses of vowel merger (see Section 2.3.2.2.1), one wonders if doing so would create a false representation of the data in cases where the vowels actually occur in unevenly distributed phonetic contexts (i.e., in cases where remnants of allophonic conditioning can be seen in the lexicon despite the vowels currently being contrastive, such that an imbalance of phonetic contexts in the data is reflective of actual tendencies in the lexicon rather than of imbalances in a particular data set). In other words, the results of the present study, together with Evanini's discussion of the English low-back merger, suggest that there are identifiable reasons why vowels that have progressed toward a more merged state might never reach a point of being acoustically indistinguishable, calling into question the idea that near-mergers (in which speakers maintain small differences in production that they and/or other members of their community do not perceive) are generally indicative of a sound change having not reached "completion" at the individual or community level. Both the results of the present study and Evanini's discussion suggest that a complete lack of measurable acoustic differences is not a necessary condition for a sound change to be considered complete, and that factors such as local perception and distribution of phonetic contexts must be considered in evaluating whether merger has been reached at the individual or community level.

Chapter 7 Conclusion

7.1 Chapter summary

This chapter discusses implications of the results of various analyses for the broad research questions outlined in Chapter 1– namely, 1) effects of social and phonological factors on production, perception, and production-perception links, and 2) implications of the results for how distinctiveness and "merger" of vowels should be defined. The most important takeaway related to the first question is that social factors (social salience of vowels and social information in the task) seem to be able to facilitate perception of allophonic differences, allowing listeners to successfully categorize variants that they did not necessarily discriminate in the AX task. Regarding the second question, the results of this study suggest that context-independent definitions of concepts such as "merger" and "near-merger" should be avoided. Other contributions of the dissertation are then discussed, including its role in documenting the contemporary state of regional Armenian dialects, and various lessons about methodology that emerged from the analyses. Finally, possible directions for future investigations of both production and perception are presented.

7.2 Implications for broad research questions

7.2.1 Influence of social and phonological factors on perception and production

The first overarching research question introduced in Chapter 1 asked how social and phonological factors influence production and perception of vowels, as well as relationships between production and perception. Social factors explored in the analyses included demographic characteristics of speakers/listeners, social salience of linguistic features, and social information in the task, while phonological factors included the differing natures (e.g., allophonic vs. contrastive) of the vowel differences explored in the dissertation. Social variation was found in the production of some vowels, particularly in relation to gender, and larger differences were generally produced for the socially salient allophonic [a]/[ɒ]$_{A}$ vowel pair than for the contrastive but less salient [ɔ]/[ɒ]$_{O}$ vowel pair. Evidence for the influence of both

phonological and social factors on perceptual outcomes was found, as participants discriminated the allophonic [a]/[ɒ]$_{A}$ vowel pair with difficulty in the AX task, but categorized it more accurately in the socially-framed dialect categorization task, likely thanks to its social salience. Results related to production-perception links were very limited and difficult to interpret, with the most interesting finding being that men categorized vowels less accurately in the dialect categorization task (suggesting lower metalinguistic awareness), despite using some dialect vowels more than women in production.

7.2.1.1 Production

The investigation of social variation in the maintenance of dialect vowels in the 2025 production data formed an extension of the analysis of the 2022 data (presented in Portugal [2025]). Both production studies explored relationships between various demographic factors and maintenance of the Bayazet dialect's vowel system in Gavaṛ (GENDER, EDUCATION LEVEL, and BIRTH YEAR in the 2022 analysis, and GENDER, EDUCATION LEVEL, BIRTH YEAR, PREVIOUS, and HATSARAT in the 2025 analysis).

Although the 2025 analysis included only a subset of the vowels included in the 2022 analysis, the two analyses had somewhat similar findings, as GENDER was the demographic factor that was most consistently relevant for predicting variation in both analyses. Gender-based variation was found for three vowel pairs in the 2022 analysis ([a]/[ɒ]$_{A}$, [ɔ]/[ɒ]$_{O}$, and [ɒ]$_{A}$/[ɒ]$_{O}$) and for two vowel pairs in the 2025 analysis ([a]/[ɒ]$_{A}$ and [ɒ]$_{A}$/[ɒ]$_{O}$), with men maintaining dialect vowels to a greater degree than women in all cases. In the 2025 analysis, being a previous (i.e., repeat) participant also turned out to be crucial for explaining variation and explaining the differences between the 2022 and 2025 results. More specifically, examination of the interaction between GENDER and PREVIOUS for the [ɔ]/[ɒ]$_{O}$ vowel pair in the 2025 data showed that the previous participants exhibited the same type of gender pattern found in the 2022 data, while the new participants did not. This discrepancy demonstrated the potential implications of the participant sampling methods used, which are discussed in Section 7.3 below along with other methodological implications of the comparison of the 2022 and 2025 production analyses.

Demographic factors other than GENDER were sometimes relevant in supplemental analyses of the 2022 data. In the small analysis of the [a]/[ɒ]$_{A}$ vowel pair in spontaneous speech (see Portugal 2025), GENDER and BIRTH YEAR were both significant predictors, with men and older people using more dialect-like pronunciations. In the analysis of the vowel [æ] (seemingly

undergoing merger by transfer for some speakers) in the 2022 data, both GENDER and EDUCATION LEVEL were significant, with men and people with secondary education having larger distances between their [æ] and [a] categories (i.e., maintaining a more dialect-like vowel system). Thus, while there is some evidence for the relevance of other social factors in explaining variation in the maintenance of dialect vowels, GENDER was the factor with the most consistent relevance across all analyses.

Regarding how phonological factors relate to the production results, the most notable finding is the fact that the acoustic differences produced by participants for the allophonic [a]/[ɒ]$_A$ pair were overall considerably larger than those produced for the contrastive [ɔ]/[ɒ]$_O$ pair (see discussion in Section 6.4.2.3). It seems possible that the social salience of [a]/[ɒ]$_A$ as a dialect feature could be related to this discrepancy. In other words, the larger acoustic differences produced for this salient vowel pair could be related to its being used more consistently to convey social meaning. Such an interpretation is also in line with the social variation found in the production results, as the gender pattern found in relation to this vowel pair was more consistent across the 2022 and 2025 data sets in comparison with the findings in relation to [ɔ]/[ɒ]$_O$. On the other hand, Seyfarth and Garellek's (2018) finding that small, potentially coarticulatory differences from [a] (that are similar in directionality to the larger differences found in Gavaṛ's dialect) are used by speakers in Yerevan indicates that we should not necessarily assume that any potential dialect leveling in Gavaṛ would result in complete overlap of [a] and [ɒ]$_A$. In other words, if some speakers in Gavaṛ are converging toward the vowel system of the supralocally influential Yerevan dialect, we would still expect them to maintain some difference between [a] and [ɒ]$_A$ based on Seyfarth and Garellek's finding. Thus, the maintenance of larger differences in the allophonic [a]/[ɒ]$_A$ pair than in the contrastive [ɔ]/[ɒ]$_O$ pair could be related not only to social salience as a dialect feature, but also to the phonological characteristics of the dominant supralocal variety, which may facilitate the maintenance of differences in [a]/[ɒ]$_A$.

7.2.1.2 Perception

7.2.1.2.1 AX task – allophonic vowels discriminated poorly despite large acoustic differences

The main result of the analysis of the AX task presented in Chapter 4 was relatively poor discrimination of the [a]/[ɒ]$_A$ vowel pair. The acoustic differences presented in the stimuli were

much larger for this vowel pair than for [ɔ]/[ɒ]$_O$, such that discrimination of [a]/[ɒ]$_A$ would have been expected to be better than for [ɔ]/[ɒ]$_O$ if size of acoustic differences were the main factor influencing accuracy of discrimination. Nevertheless, discrimination of the two vowel pairs was very similar according to several metrics (ACCURACY, CONSISTENT INTERVALS ACCURACY, *d'*). While the size of acoustic differences was relevant for accuracy of discrimination within each vowel pair, such that the stimuli word with larger differences was discriminated more accurately for both vowel pairs, the size of acoustic differences in the stimuli alone could not explain the similar performance across the two vowel pairs.

The unexpectedly poor discrimination of [a]/[ɒ]$_A$ can likely be explained by the allophonic nature of this vowel pair, as allophonic differences have generally been found to be perceived less accurately than contrastive ones (see literature review in Section 3.4.2). As noted in Chapter 4, discrimination of [a]/[ɒ]$_A$ was particularly poor toward the end of the vowel continuum representing small differences from the SEA-like pronunciation [a], such that no participants consistently discriminated the 1-step or 2-step intervals near this end of the continuum. This pattern was consistent across both stimuli words representing [a]/[ɒ]$_A$. The especially poor discrimination toward the SEA-like end of the continuum for this vowel pair has several possible explanations. As noted in Chapter 4, small differences toward this end of the [a]/[ɒ]$_A$ continuum are most analogous to the small, potentially coarticulatory differences identified in Yerevan by Seyfarth and Garellek (2018). Accordingly, it is possible that listeners perceived these intervals as small coarticulatory differences rather than as the Bayazet dialect's characteristic (large) allophonic differences, thus attributing them to their consonantal environment and having difficulty discriminating them. The "perceptual magnet effect" (see Kuhl 1991), in which listeners have trouble discriminating small differences from a category prototype, could also explain the response pattern. However, since the same type of response pattern was not seen for the [ɔ]/[ɒ]$_O$ vowel pair, for which discrimination was slightly or substantially better near the SEA-like end (depending on the stimuli word), it seems unlikely that the perceptual magnet effect is the only explanation for the pattern of results found for [a]/[ɒ]$_A$.

7.2.1.2.2 Dialect categorization task – social salience of vowels facilitates categorization

According to the analysis of the dialect categorization task presented in Chapter 5, participants responded to the [a]/[ɒ]$_A$ vowel pair in a more expected manner than the [ɔ]/[ɒ]$_O$ vowel pair. This result emerged both from the analysis of the pooled data using mixed-effects

logistic regression models and from the analysis of metrics calculated at the individual level. In the pooled mixed-effects logistic regression models, the words representing [a]/[ɒ]$_A$ had more significantly negative slopes, reflecting the less-noisy data visualizations of this vowel pair and aligning more closely with the expected response pattern of having more "Gavaṛ" responses at the dialect-like end of the continuum and more "not Gavaṛ" responses at the SEA-like end of the continuum. According to the analysis of individual-level metrics, the [a]/[ɒ]$_A$ vowel pair had more participants exhibiting the expected response pattern, and responses to this vowel pair were also generally steeper (i.e., more categorical), aligning with the results of the group-level models in suggesting that participants responded to this vowel pair in a more expected way. Although participants exhibited a great deal of individual variation in their responses to the various stimuli words, it is notable that the single most common overall response pattern (exhibited by 11 participants) was to have the expected pattern for all three of the words representing [a]/[ɒ]$_A$ and none of the words representing [ɔ]/[ɒ]$_O$. As [a]/[ɒ]$_A$ is more salient than [ɔ]/[ɒ]$_O$ as a dialect feature according both to previous literature (e.g., Katvalyan 2016) and to the 2022 sociolinguistic interviews (see Chapter 2), social salience is the most likely explanation for the discrepancy in categorization between the two vowel pairs.

7.2.1.2.3 Comparison of two perception tasks – broad differences between vowel pairs and individual differences in performance across the two tasks

As demonstrated by the results discussed in the previous two sections, the allophonic [a]/[ɒ]$_A$ vowel pair was discriminated less accurately in comparison with [ɔ]/[ɒ]$_O$ than would have been expected based on the size of acoustic differences presented in the stimuli. In contrast, [a]/[ɒ]$_A$ was categorized more accurately in the dialect categorization task. Thus, the allophonic nature of [a]/[ɒ]$_A$ appears to have inhibited its discrimination, while its social salience as a dialect feature may have facilitated its categorization in a task with explicit social framing. On the other hand, [ɔ]/[ɒ]$_O$ being contrastive and likely less salient could explain why it was discriminated more accurately (considering the size of acoustic differences presented), but categorized less accurately. Overall, comparison of the two tasks thus demonstrates how both social factors (e.g., social salience) and linguistic factors (e.g., a sound's phonological status) can influence perceptual outcomes, highlighting how social factors can facilitate perception of allophonic and/or coarticulatory differences.

It is also notable that the locations of individual perceptual boundaries calculated from responses to the dialect categorization task differed across the two vowel pairs in a way that was consistent with the results of the AX task. For [a]/[ɒ]$_A$, these boundaries were on average near the middle of the 7-step continuum, while for [ɔ]/[ɒ]$_O$, they were (significantly or sub-significantly, depending on the model) closer to the right (i.e., SEA-like) end of the continuum. Similarly, as noted above, discrimination was poor near the SEA-like end of the continuum for [a]/[ɒ]$_A$ but not for [ɔ]/[ɒ]$_O$. Comparison of these two results suggests that small differences from the SEA-like variant [a], which participants had difficulty discriminating in the AX task, are also not generally the differences that triggered dialect association in the categorization task. Rather, the average boundary location being near the middle of the continuum suggests that it was variants more dramatically different from [a] that were generally perceived as dialect variants. This suggests that small differences from [a] are perceived as the same type of low-level allophonic or coarticulatory phenomena found in Yerevan Armenian, while larger differences from [a] are perceived as Gavaṛ's characteristic dialectal allophone. Though it is important to consider the caveat that the statistical analysis of BOUNDARY LOCATION included participants whose response patterns were in both the expected and opposite directions (meaning that some participants included in this analysis categorized the SEA-like end of the continuum as dialectal), such participants were relatively few for the [a]/[ɒ]$_A$ vowel pair (see Figure 5.16), and the figure presented in Appendix F suggests that such participants' boundary locations were generally not outliers that would have skewed the results of the quantitative analysis. In any case, the overall tendency for boundary locations to not fall in areas of poor AX discrimination holds regardless of the existence of such participants.

As noted in Chapter 6, the results of quantitative comparisons of performance on the two tasks were limited, with any effects found in various models being weak and sometimes contradictory. According to some models, participants with higher AX accuracy (as measured by *d'*) for [a]/[ɒ]$_A$ also had more expected and/or more categorical responses for this vowel in the dialect categorization task, but this effect did not always reach significance, and an opposite (not reaching significance) tendency was found in relation to a different metric of AX performance (ACCURACY).

The inconclusive results of the quantitative comparison of the two tasks are unsurprising given the degree of individual variation uncovered in the qualitative examination of performance

across the two tasks (presented in Chapter 6). Various sub-categories of participants could be identified – for example, some participants had difficulty categorizing stimuli in the dialect categorization task despite good AX discrimination, while others showed an opposite tendency. For some participants, the location of the perceptual boundary in the AX task was transparently related to AX performance in some way, while for others, there was no obvious relationship. Thus, while group-level analyses pointed to a tendency for the dialect categorization task to activate sensitivity to small acoustic differences that were not discriminated in the AX task (particularly for the allophonic and socially salient [a]/[ɒ]$_A$ pair), the individual-level analysis revealed a great deal of variation, in which relationships between responses to the two tasks differed for different participants. Comparison of the two vowel pairs in this analysis was generally uninformative (due to variation among the participants being so vast); however, it was apparent that [a]/[ɒ]$_A$ (represented by the words *mas* and *pat*) had a relatively large number of participants whose BOUNDARY LOCATION in the dialect categorization task corresponded with their SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED from the AX task. Though it is important to avoid overinterpreting this tendency, it could suggest that for the given participants, AX discrimination of [a]/[ɒ]$_A$ was possible only for those differences that were also perceived as Gavaṛ's social salient dialectal allophony. This result is also consistent with (and, in fact, is possibly a byproduct of) the poor discrimination of [a]/[ɒ]$_A$ near the SEA-like end of the continuum, which necessitates participants' SMALLEST INTERVAL CONSISTENTLY DISCRIMINATED lying closer to the middle of the continuum, which is also where perceptual boundaries were located on average for this vowel pair.

7.2.1.2.4 Summary of overarching contributions related to perception

In general, the results of the two-part perception experiment suggest that both phonological and social factors can influence perception of vowels. While the results of the AX task aligned with previous literature in suggesting that discrimination of vowels is influenced by their phonological status (i.e., allophonic and/or coarticulatory vs. contrastive), the results of the dialect categorization task suggest that social factors (such as salience of the given vowel and social prompting in the task) can facilitate perception of allophonic differences, allowing participants to attend to and use small differences that they discriminated with difficulty in the AX task. As discussed in Chapter 3, though previous research about the effects of social factors on speech perception is plentiful, the effects of social factors on perception of non-contrastive

phonetic detail are less well-studied, especially in the case of perception of coarticulatory effects. This dissertation's findings related to how social factors might influence perception of such phenomena thus form an important contribution to the literature about the effects of social factors on speech perception.

Though previous studies have also explored the implications of activating different perceptual behaviors using different types of tasks, generally contrasting acoustic discrimination and semantic categorization of minimal pairs, the perception experiment presented in this dissertation departs from previous literature by comparing acoustic discrimination with evaluation of dialect association (as operationalized through categorization of non-contrastive variants). As discussed in Chapter 3, the choice to compare these two types of tasks was largely motivated by the characteristics of the allophonic [a]/[ɒ]$_A$ pair, as it was hypothesized that it could function as / be perceived as a synchronic coarticulatory phenomenon or a socially salient dialectal phenomenon depending on the task type and on the size of the acoustic differences presented. In other words, it was hypothesized that different types of tasks could activate different ways of perceiving this ambiguous phenomenon, and the results (as described above) were generally consistent with this hypothesis.

7.2.1.3 Production-perception links

7.2.1.3.1 Uninformative quantitative analyses

Like the quantitative comparison of the two perception tasks, the quantitative comparison of perception and production was relatively uninformative, providing little evidence related to the existence of perception-production links or to how they might be affected by social and phonological factors. No relationships between production and perception were found in any analyses of the AX task. Some significant and sub-significant effects were found in analyses of the dialect categorization task, but these are difficult to interpret either because of the nature of the models in which they were found, or because of the methods on which the analyses were based. In particular, relationships (one significant, one sub-significant) between production metrics (ED-median or Pillai scores) and responses were found for both vowel pairs in the group-level logistic regression models, such that participants with more dialect-like pronunciations chose "Gavaṛ" significantly less for [a]/[ɒ]$_A$ and sub-significantly more for [ɔ]/[ɒ]$_O$. It is unclear why such a discrepancy between the two vowel pairs would exist, and it is

also unclear how informative these relationships are in general, since these models explored relationships between production metrics and overall "Gavaṛ" vs. "not Gavaṛ" responses, rather than more specific aspects of response patterns.

In the models predicting individual-level metrics, a single sub-significant relationship between production and perception was found. In particular, for [a]/[ɒ]$_A$, participants with more dialect-like pronunciations had perceptual boundaries further toward the dialect-like end of the continuum. This result could suggest an interesting link between producing more dialect-like variants and only perceiving relatively extreme dialect-like pronunciations as dialectal. The trend from the group-level analysis, in which participants with more dialect-like pronunciations of the salient [a]/[ɒ]$_A$ pair chose "Gavaṛ" less overall when categorizing this pair, could also be a manifestation of the same tendency. Nevertheless, it is important to remember that the analysis of BOUNDARY LOCATION may not have been ideal given the nature of the task (see discussion in Section 5.6.1), and also that the models predicting BOUNDARY LOCATION included all participants with categorical response patterns (i.e., both expected and opposite – see Section 5.7.2.2), complicating the interpretation of the results. Even so, as the figures presented in Appendix F suggest, the link between dialect-like pronunciations and perception boundaries located further toward the dialect-like end of the continuum was driven primarily by participants with the expected response pattern, which provides support for the aforementioned production-perception link. An alternative, more suitable method of determining boundary locations (e.g., qualitatively, through visual examination of participants' responses) likely could have been created if responses to the dialect categorization task were generally less noisy. However, due to the vast variety of response patterns obtained, it was not possible to come up with a method that would not have been extremely subjective.

As discussed in Section 6.5.2, the general lack of clear perception-production links could be related to the need for listeners to understand the variety of speech forms found in their community, regardless of whether they themselves produce those forms. Such an explanation for weak or lacking relationships between production and perception has been noted by many previous authors (see Section 3.4), including Beddor (2015).

7.2.1.3.2 Further evidence for poor discrimination of allophonic vowels

Despite methodological limitations (see Section 6.4.2.4), the additional qualitative analysis of production-perception relationships was somewhat more informative than the

quantitative analysis. The results of the qualitative analysis aligned with the overall results of the AX task in suggesting that discrimination of the allophonic [a]/[ɒ]$_A$ pair was unexpectedly poor, considering the large acoustic differences between the members of this pair both in the stimuli and in most participants' own productions. In addition to comparing each participant's productions with their own perception and with the perception of the best discriminators in the sample, this analysis compared each participant's production and perception with several thresholds for discriminability, merger, or near-merger used in previous literature. According to all explored methods the number of participants who produced discriminable/unmerged vowels was greater for [a]/[ɒ]$_A$ than for [ɔ]/[ɒ]$_O$, often substantially so. In spite of this, most participants would have been unlikely to discriminate their own produced differences even for [a]/[ɒ]$_A$ (for which most participants produced relatively large differences), and around two-thirds of participants' produced differences in this vowel pair would likely not have been discriminated by anyone in the sample, despite being discriminable according to methods from previous literature. Thus, the results demonstrate that not all differences deemed discriminable by methods found in previous literature will actually be consistently discriminated in any given speech community, and that factors such as phonology of the given language variety (e.g., allophonic nature of [a]/[ɒ]$_A$) likely affect discrimination of small vowel differences.

7.2.1.3.3 Broad demographic patterns – different social factors are relevant for production and perception; production and perception not straightforwardly related

Broad comparison of the production and perception results showed that different demographic factors were relevant for explaining variation in production and perception. As noted above, GENDER was the most consistently relevant demographic factor in the production analyses. BIRTH YEAR was found to be relevant for explaining variation in responses to both perception tasks, with older people discriminating stimuli more accurately in the AX task and categorizing them more accurately (i.e., in a more expected manner) in the dialect categorization task. EDUCATION LEVEL was also relevant for the AX task, with people without higher education discriminating stimuli more accurately, while GENDER was relevant in the dialect categorization task, with women categorizing stimuli in a more expected manner than men. It is notable that out of the two perception tasks, the dialect categorization task was the one where GENDER was found to be relevant, as this suggests that social variation in the responses to this task was more related to production in comparison with social variation in the AX task. However, while men used

more dialect-like vowels than women in production, they were worse than women at recognizing as dialectal and categorizing these vowels in perception. This suggests that using a feature in one's own speech is not necessarily related to perception of that feature, which aligns with Freeman's (2023: 16) finding according to which younger participants produced vowel differences that they had difficulty perceiving, while older participants perceived differences that they did not necessarily produce. While mismatches related to age can be indicative of sound change in progress (in the case of Freeman's study, the reversal of a merger; see also Kuang and Cui [2018] regarding sound change and age-related production-perception mismatches), it is less clear what a gender-related mismatch might indicate. As discussed in Chapter 6, it is possible that it simply reflects women's tendency to adopt "change from above" in production more quickly than men, while also reflecting the increased metalinguistic awareness that could compel them to adopt such changes in production.

Overall, the most important takeaway from the comparison of broad demographic patterns in the production and perception data is that production and perception are not necessarily related in a straightforward manner. This finding is consistent with the general lack of clear findings in the quantitative analyses comparing production and perception. The idea that those who produce a certain feature might generally have more difficulty perceiving it is also supported by a number of studies reviewed by Baese-Berk et al. (2025), which find that participants learning sounds/contrasts from unfamiliar languages perceive those new sounds more accurately when they have received only perceptual training, in comparison with participants who have also been asked to produce the given sounds while learning them.

7.2.2 Defining vowel distinction / implications for concepts of "merger" and "near-merger"

The second broad research question of this dissertation explores what it means for vowel sounds to be distinct from one another in production and perception, with consideration for the fact that social and phonological factors can influence both production and perception. As noted in Chapter 1, although these questions have implications for definitions of concepts such as "merger" and "near-merger," the goal of this dissertation is not to comment on what it means for vowels to be contrastive. Instead, the goal is to explore how perception of vowels with various phonological and social characteristics relates to the size/measurability of acoustic differences in production between those vowels, along with the implications of such relationships.

As discussed in Chapters 1 and 3, defining merger in both production and perception is more complex than it might initially appear. There is no consensus about how merger should be defined acoustically (i.e., about what size of acoustic difference is small enough to be considered negligible), and perception of mergedness of vowels has been found to be affected by factors such as social information about the talker. Thus, distinguishing merger (when there is no difference between vowels in both production and perception) from near-merger (where a small difference is maintained in production, which is not generally perceived in the given speech community) is no simple task. A careful examination of previous literature reveals that the line between near-merger and allophonic variation (and, more generally, non-contrastive phonetic detail that expresses social meaning, regional affiliation, etc.) is also somewhat blurry, despite near-merger being a label that is generally applied in situations where contrastive vowels are merging. In discussing how it might be possible for speakers to consistently produce distinctions that they do not perceive (i.e., near-merger), several authors (Herold 1990: 214-215; Janson & Schulman 1983: 334; Labov 1972: 254) invoke explanations that draw on speakers' abilities to subconsciously acquire allophonic patterns and dialect-specific phonetic detail. Janson and Schulman's (1983: 334) discussion of a case of near-merger draws a particularly direct connection between near-merger, allophony, and the expression of social meaning, as they propose that the near-merger that they identify is used (similarly to allophony) to convey information about the speakers' dialect or social characteristics, rather than to convey semantic information. Thus, exploring the production and perception of allophony (including the role of social meaning) is not as irrelevant to discussions of merger and near-merger as it might initially appear. As noted in Evanini's (2009: 141-142) discussion of the English low-back vowel merger, even contrastive vowel differences can show remnants of historical allophonic conditioning, such that small consistent phonetic differences might persist in production even when speakers are merged in perception. Thus, the study of perception and production of allophonic vowel differences is also relevant for understanding vowel mergers because vowels that are contrastive with one another might nevertheless show an uneven distribution of surrounding phonetic contexts in the lexicon.

As noted above, the results of this dissertation suggest that discriminability of vowels is indeed influenced by factors other than the size of acoustic differences, as the allophonic [a]/[ɒ]$_A$ vowel pair was not discriminated with higher accuracy in the AX task despite having

substantially larger differences in its stimuli than the [ɔ]/[ɒ]$_O$ pair. While the phonological characteristics of the two vowel pairs thus seem to have influenced their discrimination in the AX task, social factors (e.g., salience of [a]/[ɒ]$_A$) seem to have played a larger role in the social categorization task, where [a]/[ɒ]$_A$ was categorized in a more expected manner despite its relatively poor discrimination in the AX task. As the qualitative individual analysis presented in Chapter 6 showed, this group-level tendency was clearly reflected in the response patterns of some individual participants, as there were some participants who discriminated this vowel pair very poorly in the AX task, but nevertheless categorized it clearly in the dialect categorization task. Thus, the social framing of the dialect categorization task seems to have allowed some participants to categorize as different stimuli that they did not discriminate in the AX task. The results of the qualitative production-perception analysis presented in Chapter 6 also aligned with the general results of the AX task in suggesting that participants' generally had difficulty discriminating differences in [a]/[ɒ]$_A$ that were relatively large, and which would have been deemed likely to be discriminable by various methods used in previous literature. Overall, the results thus suggest that the size of vowel differences that listeners can discriminate and/or categorize as different depends on factors such as social and phonological characteristics of the vowels themselves, as well as on the task type.

Accordingly, it seems that distinctiveness or merger of vowels cannot be defined in any context-independent way. As alluded to by the discussion above, there may be certain identifiable factors (other than allophony) that cause small phonetic differences between vowels to persist in spite of a lack of perception of these differences. These include remnants of historically existing allophony (as noted by Evanini [2009]) and use of the small vowel differences to convey social meaning such as dialect association, rather than semantic meaning (as noted by Janson & Schulman [1983]), and perhaps also other factors that have yet to be identified. It is unclear to what extent such factors should prevent vowels from being considered "merged" (and, conversely, should cause them to be considered "nearly merged"), since the definition of near-merger is in general conceptually and quantitatively unclear (see discussion in Chapter 3). In particular, it is not clear if this term should be applied in cases where there is an identifiable reason for the maintenance of consistent small differences in production, or if it should be reserved for cases where no such explanation can be identified. In any case, the

practice sometimes found in previous literature of defining acoustic thresholds for merger and near-merger seems to be empirically unsupported.

7.3 Methodological contributions

7.3.1 Acoustic analysis – potential implications of Bark conversion

As discussed in Chapter 2, the question of whether or not vowel formant data should be converted from Hz to Bark before being subjected to further analyses has not always been treated with care in previous sociophonetic studies. Though there is good general justification for converting to Bark, particularly in studies where perception of vowels is being explored (as opposed to only production), previous studies do not always provide an explanation of the choice to convert or not convert formant data, nor do they generally explore the implications of converting or not converting the data. Though various authors have converted data to Bark before calculating Pillai scores, the analysis of Bark-converted and non-Bark-converted data in this study suggested that Bark-conversion does not affect Pillai scores. On the other hand, it does affect calculations of Euclidean distance, and in one case the demographic factors that were found to be significant predictors of Euclidean distance were different in the analyses using Bark-converted and non-Bark-converted data (see Section 2.5.1.2).

Ultimately, it seems that the decision to convert formant data to Bark is based on principle – i.e., it is justified by the Bark scale's goal of bringing acoustic formant data closer to the reality of how it is processed by the human auditory system. However, at least when studying production and social variation in production, it is difficult to demonstrate that Bark-converted formant data is capturing variation in a more real or more meaningful way than non-Bark-converted data. In other words, since we have no way to postulate the true nature of social variation in the production of vowels independently of various measurement and data processing techniques, it is not possible to say whether Bark-converted formant data is a more or less accurate reflection of reality (especially given that production and perception of vowels often seem to not be related in a straightforward way – see literature review in Chapter 3). Nevertheless, given that the decision to convert data to Bark can have implications for the results of further analyses (particularly those involving metrics like Euclidean distance), it would be ideal for the field of sociophonetics either to adopt a single convention (i.e., to always convert to

Bark or always not convert to Bark), or for studies to discuss the results of both types of analyses.

7.3.2 Methodological implications of comparison of production analyses – participant sampling, token sample size, recording quality, elicitation methods

As discussed in Chapter 2, the results of the 2022 and 2025 production analyses differed in several ways that likely resulted from various methodological decisions. Thus, comparison of the two analyses highlights the potential impacts of different methods, underscoring the importance of weighing the advantages and disadvantages of certain decisions when designing a study.

Firstly, comparison of the results for the [ɔ]/[ɒ]o vowel pair in 2022 and 2025 demonstrated how issues related to participant sampling can affect which social patterns are found or not found in a given data set. For this vowel pair, a gender pattern was found in 2022 but not 2025. However, examination of the 2025 data of only those participants who had also participated in 2022 revealed that this subset of participants exhibited a similar gender pattern in both data sets. Accordingly, the discrepancy between the 2022 and 2025 data sets likely simply reflects the differing characteristics of the two participant samples. As neither sample was random, the degree to which social patterns can be generalized to the population of Gavaṛ is already limited, and it is difficult to say which sample might be a more accurate reflection of social variation in the broader community in relation to the given vowel pair. It is possible that the 2022 sample could be less representative of the population than the 2025 sample due to the 2022 participants potentially being more tightly connected within a social network. However, it is also possible that one or both samples were simply too small to be representative regardless of the sampling method, especially given the degree of linguistic heterogeneity that exists in Armenia. Additionally, obtaining a truly random sample (in which no participants had any connection to one another) might also have been prohibitively difficult due to the city's small population, and it is questionable if such a sample could even be considered truly random or representative due to the artificial nature of prohibiting connections between members of a small and densely socially-connected population. In any case, as there is evidence that the direct social-network sampling used in 2022 contributed to the elicitation of more natural/colloquial speech, the discrepancy between the results of the two studies highlights the need to weigh the

advantages and disadvantages of different sampling methods, with consideration for the goals of the study and for the context in which one is working.

Differences between the 2022 and 2025 data sets also included the actual Pillai scores and ED-median values for some vowel pairs. For example, for the [ɔ]/[ɒ]$_O$ vowel pair, both metrics had substantially smaller values in 2025. As discussed in Chapter 2, the fact that the thresholds calculated with the method of Stanley and Sneller (2023) divided the Pillai scores into merged and unmerged groups in a similar manner in both data sets suggests that the discrepancy in the values of the metrics is largely due to differing sample sizes. For the given vowel pair, the sample size was much larger in 2025, which is known to generally lead to smaller Pillai scores (see Stanley & Sneller 2023). It is possible that the smaller sample size in 2022 affected Euclidean distance calculations in a similar manner, as measures of central tendency of the different vowel classes (e.g., median) would have been less certain than they were when calculated from the larger sample size of the 2025 data. Both metrics could also have been affected by the poorer recording quality of the 2022 data, which also would have led to increased uncertainty in the location of formants and thus increased uncertainty in measures of central tendency. Overall, the discrepancy between the two data sets and the apparent influence of factors like sample size and recording quality demonstrate the importance of being cautious when comparing different data sets, as the values of metrics calculated from different data sets may differ for various reasons related to the data collection methods.

Though the group of words where the vowel [æ] was expected was not a main component of the 2025 analysis, comparison of the productions of this group of words in the 2022 and 2025 words lists revealed that elicitation methods also could have contributed to differences between the two data sets. As the figures presented in Chapter 2 showed, pronunciations of the [æ] vowel were relatively common in the 2022 word list but were completely absent for most participants in 2025. This included several repeat participants who had numerous pronunciations of [æ] in 2022 but pronounced all relevant words with [a] in 2025. This discrepancy suggests that the differing elicitation methods used in the two studies affected the degree to which colloquial/dialectal speech was successfully elicited. In particular, as the 2022 word list was embedded within a larger interview, immediately followed a picture-naming task, and was recorded in environments that were familiar for the participants, it seems to have been more successful at eliciting dialect-like pronunciations (at least for the [æ] vowel, which seems to be

undergoing a merger by transfer, unlike other dialect vowels explored in the production analyses). Comparison of the two data sets thus demonstrates that even when the general style of elicited speech (e.g., spontaneous speech vs. read speech) is the same in two data sets, more subtle aspects of the elicitation methods can influence results.

7.3.3 Implications of choices in data processing and statistical analyses

In many of the analyses presented in this dissertation, several different methodological choices could have been made, without there being a particularly strong reason for choosing one method over the other. For example, in the analyses of individual-level metrics of performance on the perception tasks (presented in Chapters 4 and 5), it was possible to run linear and logistic regression models that either pooled the data from the two vowel pairs or separated them into different models. The pooled models had the advantage of being able to perform direct quantitative comparison of the two vowel pairs, while they also had the disadvantage of suffering from more collinearity issues than the separated models. As there were thus arguments for choosing both options, both pooled and separate models were run, and their results were discussed and compared. In many cases, different demographic predictors would be significant in the separate models for the two vowel pairs, but these predictors' interactions with VOWEL PAIR in the pooled models would not reach significance, even if similar tendencies (in terms of directionality of effects) were observed in both sets of models. As such, if the separate models had not been run, the effects found in the separate models might never have been noticed, and the methodological decision of only running pooled models would have had a substantial effect on the interpretation of the data. This issue underscores the need for caution in interpretation of statistical significance (or lack thereof), transparency about potential effects of methodological decisions on statistical results, and exploratory analyses in cases where there is no strong argument for choosing one type of analysis over another.

A similar need to explore and compare different methods emerged from the analysis of the AX task, where different metrics of quantifying overall performance on the task were explored. Specifically, this analysis compared *d'* (which incorporates both correct responses to "different" trials and incorrect responses to "same" trials) with the proportion of correct responses on "different" trials only (ACCURACY). While the differences in the results for the different words and vowel pairs appeared overall similar for the two metrics based on data visualizations, the statistical significance of differences between the words and vowel pairs was

not identical across the two metrics. As discussed in Chapter 5, the two metrics also sometimes appeared to relate to performance on the dialect categorization task in different ways. Thus, the analysis of the AX task also demonstrated that methodological decisions can have significant downstream effects, suggesting that one should compare the results of different methods and/or avoid arbitrary decisions in cases where there are reasonable arguments for multiple different methods.

7.4 Contributions to documentation of contemporary Armenian dialects

As discussed in Chapter 1, this dissertation is a continuation of the 2022 production study (presented in Portugal [2025]), which had the goal of contributing to the documentation of contemporary Armenian dialects by examining variation in the maintenance of dialect features in a regional city. The results of both the 2022 and 2025 production studies suggested that many speakers are maintaining some but not all of the dialect vowels described by Katvalyan (2016), while others are using a vowel system that is closer to SEA / the Yerevan dialect, with a great deal of individual variation and intermediate states. Thus, while the production results suggest that a "quasi-standard colloquial language" (Hodgson 2019: 95-96) has become at least somewhat established in regional cities, they also demonstrate that some dialect features are nevertheless resilient, and that a wide range of hybrid ways of speaking likely exist. Despite the limitations of the analyses of production data (i.e., relying primarily on word-list and picture-naming data), these analyses form an important contribution to knowledge about the current state of Armenian dialects, as they represent one of the first quantitative investigations of information presented in Armenian dialectological sources. The inclusion of demographically diverse participants and transparency about the participants and methods allow for a preliminary understanding of social variation in the use and maintenance of dialect features, adding much-needed nuance to the information presented in previous dialectological descriptions. As noted in Chapter 2, it is unclear whether the results of the 2022 and 2025 production studies (specifically, the finding that no speaker was consistently maintaining the entire dialect vowel system as described by Katvalyan [2016]) are indicative of ongoing sound change, or if they simply represent relatively stable variation not reflected in previous works due to the limitations of dialectological methodology. In any case, the results indicate that neither Katvalyan's (2016: 19) sweeping statements about dialect maintenance, nor other authors' broad statements about dialect leveling are likely representative of the complex linguistic reality in regional cities.

The contribution of this dissertation to the study of contemporary Armenian dialects includes not only the quantitative analysis of variation in production, but also the investigation of how previously described dialect features are perceived, and of how variation in their perception compares with variation in production. As Armenian dialectological works do not include investigations of perception, and given the interesting patterns of variation found in the 2022 production data (in particular, the fact that all participants maintained differences in the [a]/[ɒ]$_A$ vowel pair with similar directionality but vastly varying magnitude), the perception experiment sought to probe Gavaṛ residents' awareness of various types of vowel differences. In particular, the experiment sought to determine which types of produced differences were discriminable and/or socially meaningful (i.e., conveying association with Gavaṛ's dialect) vs. which types of differences would be discriminated/categorized with difficulty due to their being perceived as low-level allophonic/coarticulatory effects and/or as random noise.

In a broad sense, the results of the perception experiment were consistent with Katvalyan's descriptions of the dialect and with aspects of the 2022 and 2025 production results. For example, the unexpectedly poor discrimination of [a]/[ɒ]$_A$ aligns with its description as an allophonic difference, while the more accurate categorization of this pair in the dialect categorization task aligns with Katvalyan's (2016) description of it as the dialect vowel system's most salient feature. The apparent social salience of this vowel pair and its relatively accurate categorization in the dialect categorization task are also consistent with its production results. In particular, it showed a pattern of gender-based variation that was consistent across the two sets of production data, which suggests that it may have relatively stable and robust social meaning.

On the other hand, as discussed in Chapter 5, the results of the dialect categorization task contained a great deal of individual variation, and fewer participants than expected accurately categorized the stimuli even for the salient [a]/[ɒ]$_A$ pair. While this was likely in part due to the inherent difficulty of the task (see Section 5.7.3), it also suggests that metalinguistic awareness even of the most salient dialect features is highly variable in Gavaṛ. As discussed in Chapter 5, one participant's post-experiment commentary suggested that some of the talker's pronunciations were not extreme enough to be considered dialect speech. Based on this commentary and on the highly variable task responses, it seems possible that some participants associate the word "dialect" with villages and/or with the most extreme stereotypes or conservative speech varieties, and do not necessarily recognize less extreme urban variants as dialectal (even if they are

generally similar in nature to / exist on a clear continuum with more extreme variants). In other words, it is possible that participants who did not respond in the expected manner have some degree of metalinguistic awareness about the given dialect features, but that their awareness is based on more extreme pronunciations (stereotypically associated with villages).

As discussed in Chapters 4-6, younger people both discriminated less accurately in the AX task and categorized stimuli in a less expected manner in the dialect categorization task. As discussed in Chapter 5, these trends are consistent with the findings of a dialect classification task carried out by Dossey et al. (2020), and could be related to younger people having less lifetime exposure to dialect-like speech and thus less developed metalinguistic awareness (either due to ongoing or recent dialect leveling or simply to their age). However, there are also other possible explanations that should be considered, especially given the overall poorer-than-expected performance on the dialect categorization task. For example, perhaps metalinguistic awareness related to the concept of "dialect" is changing in Gavaṛ, such that features recognized as "dialect" are becoming more limited (i.e., perhaps more confined to extreme, stereotypically rural pronunciations), while those that are recognized as unmarked colloquial speech may be expanding to include the less extreme dialectal pronunciations found in the city. It is interesting that age-related variation was generally not found in the production analyses, as in combination with the perception results this suggests that recognition and metalinguistic awareness of dialect features may be declining, even if some of those features are relatively stable in production. This could suggest that some features (for example, the less extreme urban pronunciations of dialect vowels) are undergoing a transition from being sociolinguistic stereotypes toward something closer to "indicators" (see Labov 1972), which show social variation despite a lack of speaker awareness.

As noted in Chapter 4, it is difficult to understand why younger people would have performed worse on the AX task in particular, but this result could perhaps also be related to younger people's more limited exposure to dialectal speech over their lifetimes or to declining awareness about dialect features. In other words, if metalinguistic knowledge about dialect vowels could have helped participants to recognize and discriminate those vowels even in a task without social framing, then the relative lack and/or decline of such knowledge in younger people could have contributed to their worse discrimination in the AX task.

In any case, the vast variation found in the production data and the unexpectedly variable metalinguistic awareness reflected in the dialect categorization task together suggest that speech in Gavaṛ is characterized by variable use of dialect features that may not always be (consciously or unconsciously) recognized as dialect features. In other words, speech in Gavaṛ seems to be characterized by wide variation in the maintenance of dialect features, but many residents may not recognize the urban variants of these features as manifestations of dialectal speech. The existence of a great deal of seemingly unrecognized variation underscores the need for more documentation of the current state of Armenian dialects in regional cities. There likely exist a great deal of intermediate language forms, which are not necessarily being documented by Armenian dialectologists, but which can show us in real time how varieties of Armenian are varying and potentially changing, and how regional dialects are responding to influence from supralocal varieties and language ideologies. The lack of previous attention paid by researchers to such language forms (i.e., less extreme, conservative, or stereotypical manifestations of dialect speech, forms influenced by contact with supralocal varieties, etc.) demonstrates "erasure" (Irvine & Gal 2000: 38) of forms that are not imagined to fit neatly into the idealized version of any given variety of Armenian. While the work of Armenian dialectologists in documenting more conservative dialectal speech is undoubtedly extremely valuable, further research using a wider variety of methods and encompassing a wide variety of participants is needed in order to shed light on the complex reality that underlies oversimplified dichotomies between varieties of Armenian (e.g., SEA vs. dialect). While it is possible that some of the production data from the present study is indicative of ongoing dialect levelling (see Chapter 2), some of the social patterns found could also represent stable variation. Although contact between regional dialects and supralocal varieties has undoubtedly increased in recent decades thanks to mass media and the internet, such contact has existed at least since the codification of SEA (if not longer), such that intermediate language forms and sociolinguistic variation in the use of dialect features are likely not a new phenomenon. In considering such forms to be equally worthy of attention and documentation as SEA and more conservative dialect speech, this dissertation has taken a step toward combatting the erasure of such forms, and it is hoped that future studies will continue to shed light on the wealth of linguistic variation that exists in Armenia today.

7.5 Future directions

Both the production and perception components of this dissertation suffer from limitations that could be addressed in the future either by performing additional analyses of the existing data or by collecting additional data.

The most obvious shortcoming of the production analyses is their reliance primarily on speech that was elicited using methods that are likely to yield more monitored speech styles (namely, a word list in 2025 and a word and picture list in 2022). In order to obtain a more complete picture of sociolinguistic variation in the maintenance of dialect vowels, it would be ideal to carry out a full analysis of the conversational speech elicited in 2022. Future studies should also elicit additional data using methods that could help to further overcome the observer's paradox – for example, group interviews in which participants converse largely with one another rather than with an interviewer. In order to obtain data that is more likely to be reflective of the speech community as a whole, it would be advisable to conduct future studies using random participant sampling (though, as discussed in Chapter 2, this might reduce the naturalness/colloquialness of the obtained speech). Of course, as this dissertation focuses only on a single type of linguistic feature (vowels), it would also be necessary to analyze other features in order to gain a fuller understanding of the state of contemporary dialectal speech in Gavaṛ. One promising area for future research would be verbal morphology, as various verb forms found in Gavaṛ's dialect (such as participles and certain inflected forms – see Katvalyan [2016]) differ from SEA, and such forms were used relatively frequently in the 2022 sociolinguistic interviews.

As discussed in Chapter 6, some of the analyses of the perception experiment (in particular, the qualitative analyses of individual-level perception-production links) were limited by the structure of the experiment. In particular, the fact that the intervals of varying sizes included only differences of varying sizes from the SEA-like end of the continuum limited the conclusions that could be drawn from analysis of this portion of the data. Conducting a follow-up experiment that would include intervals of varying sizes at different points along the continuum could address some of the questions that were not able to be answered given the limitations of the experiment's structure.

It would also be productive to have people from other parts of Armenia and people without knowledge of Armenian participate in the perception experiment (at least the AX task),

as this could help to determine if Gavaṛ residents' unexpectedly poor discrimination of the allophonic [a]/[ɒ]$_A$ vowel pair differs in some way from the discrimination of speakers of other varieties of Armenian (e.g., people from Yerevan) or of people without any exposure to the sound system of Armenian. As small, phonetically conditioned (perhaps coarticulatory) differences from [a] have been found in Yerevan by Seyfarth and Garellek (2018), it is possible that Gavaṛ residents are not the only Armenian speakers who would have difficulty discriminating this vowel pair. However, perhaps the magnitude of the differences that listeners are able to discriminate would be different in Gavaṛ and Yerevan, in accordance with the differing magnitudes of the allophonic differences that are produced in the two communities, and with their differing social characteristics in the two communities. Depending on the language backgrounds of non-Armenian-speaking listeners, they could also show patterns of discrimination that differ from those of Gavaṛ residents – for example, they might discriminate [a]/[ɒ]$_A$ more easily if their languages contain a contrast in this area of the vowel space. In order to test whether Gavaṛ listeners' poor discrimination of small differences from [a] was really a result of perceiving these small differences as a coarticulatory effect, it would also be desirable to repeat the AX task with the vowels removed from their consonantal contexts (which could result in more similar AX accuracy across the continuum if compensation for coarticulation were actually playing a role in the AX results for [a]/[ɒ]$_A$). It would also be informative to examine Yerevan speakers' productions of the given vowels in a wider range of consonantal contexts than those discussed by Seyfarth & Garellek (2018), as this would provide a clearer picture of the extent to which the small differences in [a] found in Yerevan are comparable to Gavaṛ's [a]/[ɒ]$_A$ allophony.

7.6 Conclusion

In exploring the production and perception of vowels in Gavaṛ, this dissertation sought both to answer broad questions about language and to contribute to the documentation of linguistic variation in Armenia today. In exploring the potential effects of social and phonological factors on production, perception, and production-perception links, the various components of the study found evidence for the influence of both types of factors on production and perception, though findings related to production-perception links were more limited. The most notable finding was the evidence from the perception experiment for the influence of both phonological and social factors on perceptual outcomes. In particular, the socially salient,

allophonic [a]/[ɒ]$_A$ pair showed relatively poor discrimination in the AX task, especially given the large acoustic differences presented in its stimuli, but was categorized more accurately in the dialect categorization task. This discrepancy between the results of the two perception tasks suggests that social factors (such as social salience of vowels and social framing of the task) can facilitate the perception of difficult-to-discriminate allophonic differences. Additionally, the results of the perception tasks suggested that different areas of the continuum of the [a]/[ɒ]$_A$ vowel pair are perceived in different ways. In particular, the results suggested that small differences from the SEA variant [a] are possibly perceived as coarticulatory differences (and thus discriminated with the most difficulty), while larger differences might be more likely to be perceived as Gavaṛ's socially salient dialectal allophone. The dissertation also explored implications of production and perception of distinctions between vowels in Gavaṛ for concepts such as "merger" and "near-merger" of vowels, ultimately concluding that such concepts cannot be defined in any context-independent way. In documenting contemporary regional Armenian speech, variation in the use of dialect features, and Gavaṛ residents' perception of such variation, this dissertation has drawn attention to types of speech that have been relatively overlooked in previous research on Armenian, highlighting the complexities emerging from longstanding interactions between dialects and supralocal varieties.

Appendices

Appendix A: Word List

Words are presented below with the Armenian orthography, IPA transcription showing multiple possible vowel pronunciations (where applicable), and English translation. For the sake of simplicity, consonants have been transcribed with their SEA pronunciations.

1. թիվ (‘number’); [tʰiv]
2. գետին (‘ground’); [gɛtin]
3. դաշտ (‘field’); [daʃt]
4. փուշ (‘thorn’); [pʰuʃ]
5. կաթ (‘milk’); [katʰ] (SEA vowel); [kɒtʰ] (dialect vowel)
6. դառը (‘bitter’); [darə] (SEA vowel); [dærə] (dialect vowel)
7. գեշ (‘ugly’); [gɛʃ]
8. պատ (‘wall’); [pat] (SEA vowel); [pɒt] (dialect vowel)
9. տես (‘see [imperative]’); [tɛs] (SEA vowel); [ties] (dialect vowel)
10. խոտ (‘grass’); [χɔt] (SEA vowel); [χɒt] (dialect vowel)
11. գազ (‘gas’); [gaz]
12. կով (‘cow’); [kɔv]
13. տիպ (‘type’); [tip]
14. ձագ (‘baby animal’); [d͡zakʰ] (SEA vowel); [d͡zækʰ] (dialect vowel)
15. շատ (‘much, very’); [ʃat] (SEA); [ʃɒt] (dialect vowel)
16. տորթ (‘cake’); [tɔɾtʰ]
17. զուտ (‘only, purely’); [zut]
18. սոխ (‘onion’); [sɔχ] (SEA vowel); [sɒχ] (dialect vowel)
19. բաց (‘open’); [bat͡sʰ]
20. թուզ (‘fig’); [tʰuz]
21. պետ (‘commander’); [pɛt] (SEA vowel); [piet] (dialect vowel)
22. շոգ (‘hot’); [ʃɔkʰ] (SEA vowel); [ʃɒkʰ] (dialect vowel)
23. բարդ (‘complex’); [baɾtʰ]
24. քիչ (‘little’); [kʰit͡ʃʰ]
25. գող (‘thief’); [gɔʁ] (SEA vowel); [gɒʁ] (dialect vowel)
26. դար (‘century’); [daɾ]
27. բութ (‘thumb’); [butʰ]
28. կոթ (‘stem’); [kɔtʰ] (SEA vowel); [kɒtʰ] (dialect vowel)
29. խաչ (‘cross’); [χat͡ʃʰ]

30. oղ (‘ring’); [ɔʁ]
31. նավթ (‘oil’); [naftʰ] (SEA vowel); [næftʰ] (dialect vowel)
32. ափ (‘coast, shore’); [apʰ] (SEA vowel); [ɒpʰ] (dialect vowel)
33. տիq (‘tick’); [tiz]
34. խոq (‘pig’); [χɔz] (SEA vowel); [χɒz] (dialect vowel)
35. ձախ (‘left’); [d͡zaχ]
36. գողություն (‘theft’); [gɔʁutʰjun]
37. ծեփ (‘plaster’); [t͡sɛpʰ] (SEA vowel); [t͡siepʰ] (dialect vowel)
38. աջ (‘right’); [at͡ʃʰ] (SEA vowel); [ɒt͡ʃʰ] (dialect vowel)
39. թութ (‘mulberry’); [tʰutʰ]
40. փող (‘money’); [pʰɔʁ] (SEA vowel); [pʰɒʁ] (dialect vowel)
41. կետ (‘point’); [kɛt] (SEA vowel); [kiet] (dialect vowel)
42. տառ (‘letter’); [tar] (SEA vowel); [tɒr] (dialect vowel)
43. քիթ (‘nose’); [kʰitʰ]
44. օձ (‘snake’); [ɔt͡sʰ]
45. դաս (‘lesson’); [das]
46. գողեր (‘thieves’); [gɔʁɛɾ] (SEA vowel); [gɒʁieɾ] (dialect vowel)
47. փիղ (‘elephant’); [pʰiʁ]
48. մաս (‘part’); [mas] (SEA vowel); [mɒs] (dialect vowel)
49. օդ (‘air’); [ɔtʰ]
50. թաթ (‘paw’); [tʰatʰ] (SEA vowel); [tʰætʰ] (dialect vowel)
51. մարդ (‘person’); [maɾtʰ] (SEA vowel); [mɒɾtʰ] (dialect vowel)
52. ձեթ (‘vegetable oil’); [d͡zɛtʰ]
53. կապ (‘connection’); [kap] (SEA vowel); [kɒp] (dialect vowel)
54. դարակ (‘shelf’); [daɾak] (SEA vowel); [dæɾæk] (dialect vowel)
55. բոց (‘flame’); [bɔt͡sʰ] (SEA vowel); [bɒt͡sʰ] (dialect vowel)
56. պուտ (‘speck’); [put]
57. մազ (‘hair’); [maz] (SEA vowel); [mɒz] (dialect vowel)
58. գետ (‘river’); [gɛt] (SEA vowel); [giet] (dialect vowel)
59. փոս (‘hole’); [pʰɔs] (SEA vowel); [pʰɒs] (dialect vowel)
60. ցավ (‘pain’); [t͡sʰav]
61. թեք (‘tilted’); [tʰɛkʰ] (SEA vowel); [tʰiekʰ] (dialect vowel)
62. ծոց (‘bosom, embrace’); [t͡sɔt͡sʰ] (SEA vowel); [t͡sɒt͡sʰ] (dialect vowel)
63. բադ (‘duck’); [bad]
64. կես (‘half’); [kɛs]
65. նավ (‘ship’); [nav] (SEA vowel); [nɒv] (dialect vowel)
66. հեշտ (‘easy’); [hɛʃt]
67. տաս (‘ten’); [tas] (SEA vowel); [tɒs] (dialect vowel)
68. խեղճ (‘pitiable’); [χɛʁt͡ʃ] (SEA vowel); [χieʁt͡ʃ] (dialect vowel)
69. ծով (‘sea’); [t͡sɔv]
70. գաթա (‘gata [type of Armenian pastry]’); [gatʰa]
71. դարդ (‘sorrow’); [daɾd] (SEA vowel); [dæɾd] (dialect vowel)
72. աղ (‘salt’); [aʁ] (SEA vowel); [ɒʁ] (dialect vowel)
73. կով (‘cow’); [kɔv]

74. մարդ (‘person’); [maɾt^{h}] (SEA vowel); [mɒɾt^{h}] (dialect vowel)
75. փոս (‘hole’); [p^{h}ɔs] (SEA vowel); [p^{h}ɒs] (dialect vowel)
76. բադ (‘duck’); [bad]
77. թիվ (‘number’); [t^{h}iv]
78. կոթ (‘stem’); [kɔt^{h}] (SEA vowel); [kɒt^{h}] (dialect vowel)
79. խաչ (‘cross’); [χat͡ʃ^{h}]
80. գող (‘thief’); [gɔʁ] (SEA vowel); [gɒʁ] (dialect vowel)
81. տես (‘see [imperative]’); [tɛs] (SEA vowel); [ties] (dialect vowel)
82. մաս (‘part’); [mas] (SEA vowel); [mɒs] (dialect vowel)
83. խոզ (‘pig’); [χɔz] (SEA vowel); [χɒz] (dialect vowel)
84. տառ (‘letter’); [tar] (SEA vowel); [tɒr] (dialect vowel)
85. փող (‘money’); [p^{h}ɔʁ] (SEA vowel); [p^{h}ɒʁ] (dialect vowel)
86. շատ (‘much, very’); [ʃat] (SEA); [ʃɒt] (dialect vowel)
87. պետ (‘commander’); [pɛt] (SEA vowel); [piet] (dialect vowel)
88. օդ (‘air’); [ɔt^{h}]
89. մազ (‘hair’); [maz] (SEA vowel); [mɒz] (dialect vowel)
90. փուշ (‘thorn’); [p^{h}uʃ]
91. գաթա (‘gata [type of Armenian pastry]’); [gat^{h}a]
92. բոց (‘flame’); [bɔt͡s^{h}] (SEA vowel); [bɒt͡s^{h}] (dialect vowel)
93. աջ (‘right’); [at͡ʃ^{h}] (SEA vowel); [ɒt͡ʃ^{h}] (dialect vowel)
94. գողություն (‘theft’); [gɔʁut^{h}jun]
95. դար (‘century’); [daɾ]
96. խոտ (‘grass’); [χɔt] (SEA vowel); [χɒt] (dialect vowel)
97. պատ (‘wall’); [pat] (SEA vowel); [pɒt] (dialect vowel)
98. օղ (‘ring’); [ɔʁ]
99. տաս (‘ten’); [tas] (SEA vowel); [tɒs] (dialect vowel)
100. գետին (‘ground’); [gɛtin]
101. բարդ (‘complex’); [baɾt^{h}]
102. ծոց (‘bosom, embrace’); [t͡sɔt͡s^{h}] (SEA vowel); [t͡sɒt͡s^{h}] (dialect vowel)
103. գազ (‘gas’); [gaz]
104. օձ (‘snake’); [ɔt͡s^{h}]
105. դաս (‘lesson’); [das]
106. քիթ (‘nose’); [k^{h}it^{h}]
107. ծով (‘sea’); [t͡sɔv]
108. կաթ (‘milk’); [kat^{h}] (SEA vowel); [kɒt^{h}] (dialect vowel)
109. գողեր (‘thieves’); [gɔʁɛɾ] (SEA vowel); [gɒʁiɛɾ] (dialect vowel)
110. բաց (‘open’); [bat͡s^{h}]
111. քիչ (‘little’); [k^{h}it͡ʃ^{h}]
112. դար (‘century’); [daɾ]
113. գողություն (‘theft’); [gɔʁut^{h}jun]
114. շատ (‘much, very’); [ʃat] (SEA); [ʃɒt] (dialect vowel)
115. փոս (‘hole’); [p^{h}ɔs] (SEA vowel); [p^{h}ɒs] (dialect vowel)
116. աջ (‘right’); [at͡ʃ^{h}] (SEA vowel); [ɒt͡ʃ^{h}] (dialect vowel)
117. զուտ (‘only, purely’); [zut]

118. բաց (‘open’); [bat͡sʰ]
119. խոզ (‘pig’); [χɔz] (SEA vowel); [χɒz] (dialect vowel)
120. պատ (‘wall’); [pat] (SEA vowel); [pɒt] (dialect vowel)
121. օձ (‘snake’); [ɔt͡sʰ]
122. կաթ (‘milk’); [katʰ] (SEA vowel); [kɒtʰ] (dialect vowel)
123. գող (‘thief’); [gɔʁ] (SEA vowel); [gɒʁ] (dialect vowel)
124. ձեթ (‘vegetable oil’); [d͡zɛtʰ]
125. բարդ (‘complex’); [barɾtʰ]
126. ծոց (‘bosom, embrace’); [t͡sɔt͡sʰ] (SEA vowel); [t͡sɒt͡sʰ] (dialect vowel)
127. մաս (‘part’); [mas] (SEA vowel); [mɒs] (dialect vowel)
128. փող (‘money’); [pʰɔʁ] (SEA vowel); [pʰɒʁ] (dialect vowel)
129. հեշտ (‘easy’); [hɛʃt]
130. դաս (‘lesson’); [das]
131. կով (‘cow’); [kɔv]
132. տառ (‘letter’); [tar] (SEA vowel); [tɒr] (dialect vowel)
133. գողեր (‘thieves’); [gɔʁɛɾ] (SEA vowel); [gɒʁiɛɾ] (dialect vowel)
134. խաչ (‘cross’); [χat͡ʃʰ]
135. օդ (‘air’); [ɔtʰ]
136. գազ (‘gas’); [gaz]
137. բոց (‘flame’); [bɔt͡sʰ] (SEA vowel); [bɒt͡sʰ] (dialect vowel)
138. թութ (‘mulberry’); [tʰutʰ]
139. գաթա (‘gata [type of Armenian pastry]’); [gatʰa]
140. ծով (‘sea’); [t͡sɔv]
141. մազ (‘hair’); [maz] (SEA vowel); [mɒz] (dialect vowel)
142. կոթ (‘stem’); [kɔtʰ] (SEA vowel); [kɒtʰ] (dialect vowel)
143. բադ (‘duck’); [bad]
144. խոտ (‘grass’); [χɔt] (SEA vowel); [χɒt] (dialect vowel)
145. տաս (‘ten’); [tas] (SEA vowel); [tɒs] (dialect vowel)
146. կես (‘half’); [kɛs]
147. օղ (‘ring’); [ɔʁ]
148. մարդ (‘person’); [maɾtʰ] (SEA vowel); [mɒɾtʰ] (dialect vowel)
149. խոզ (‘pig’); [χɔz] (SEA vowel); [χɒz] (dialect vowel)
150. դաս (‘lesson’); [das]
151. կով (‘cow’); [kɔv]
152. պատ (‘wall’); [pat] (SEA vowel); [pɒt] (dialect vowel)
153. թուզ (‘fig’); [tʰuz]
154. օձ (‘snake’); [ɔt͡sʰ]
155. աջ (‘right’); [at͡ʃʰ] (SEA vowel); [ɒt͡ʃʰ] (dialect vowel)
156. թեք (‘tilted’); [tʰɛkʰ] (SEA vowel); [tʰiekʰ] (dialect vowel)
157. գազ (‘gas’); [gaz]
158. խոտ (‘grass’); [χɔt] (SEA vowel); [χɒt] (dialect vowel)
159. բարդ (‘complex’); [barɾtʰ]
160. գողեր (‘thieves’); [gɔʁɛɾ] (SEA vowel); [gɒʁiɛɾ] (dialect vowel)
161. մաս (‘part’); [mas] (SEA vowel); [mɒs] (dialect vowel)

162. տիq (‘tick’); [tiz]
163. բաց (‘open’); [ba͡tsʰ]
164. ծովվ (‘sea’); [t͡sɔv]
165. գաթա (‘gata [type of Armenian pastry]’); [gatʰa]
166. փոս (‘hole’); [pʰɔs] (SEA vowel); [pʰɒs] (dialect vowel)
167. բադ (‘duck’); [bad]
168. կոթ (‘stem’); [kɔtʰ] (SEA vowel); [kɒtʰ] (dialect vowel)
169. պուտ (‘speck’); [put]
170. շատ (‘much, very’); [ʃat] (SEA); [ʃɒt] (dialect vowel)
171. օղ (‘ring’); [ɔʁ]
172. մազ (‘hair’); [maz] (SEA vowel); [mɒz] (dialect vowel)
173. բոց (‘flame’); [bɔ͡tsʰ] (SEA vowel); [bɒ͡tsʰ] (dialect vowel)
174. դար (‘century’); [daɾ]
175. գողությունն (‘theft’); [gɔʁutʰjun]
176. տաս (‘ten’); [tas] (SEA vowel); [tɒs] (dialect vowel)
177. կետ (‘point’); [kɛt] (SEA vowel); [kiet] (dialect vowel)
178. փող (‘money’); [pʰɔʁ] (SEA vowel); [pʰɒʁ] (dialect vowel)
179. մարդ (‘person’); [maɾtʰ] (SEA vowel); [mɒɾtʰ] (dialect vowel)
180. օդ (‘air’); [ɔtʰ]
181. տառ (‘letter’); [tar] (SEA vowel); [tɒr] (dialect vowel)
182. ծոց (‘bosom, embrace’); [t͡sɔ͡tsʰ] (SEA vowel); [t͡sɒ͡tsʰ] (dialect vowel)
183. խաչ (‘cross’); [χat͡ʃʰ]
184. գող (‘thief’); [gɔʁ] (SEA vowel); [gɒʁ] (dialect vowel)
185. կաթ (‘milk’); [katʰ] (SEA vowel); [kɒtʰ] (dialect vowel)
186. խեղճ (‘pitiable’); [χɛʁt͡ʃ] (SEA vowel); [χiɛʁt͡ʃ] (dialect vowel)

Appendix B: Demographic and Language Background Questionnaire

Մասնակցի տվյալները

Ամսաթիվ. _____ Սեռ. _____ Ծննդյան տարեթիվը. _____ Ծննդավայր. _____ Եթե ծննդավայրից դուրս եք ապրել, նշե՛ք երբ, որտեղ և ինչքան ժամանակ. _____ _____ _____ Ներկայիս բնակության վայրը. _____	Ձեր մասնագիտությունը _____ Ներկայիս աշխատանքը. _____ Նախկինում ունեցած աշխատանքը (եթե տարբերվում է վերը նշվածներից). _____ Տանը խոսվող լեզուն/լեզուները (մինչև չափահաս դառնալը). _____ Եթե խոսում եք այլ լեզուներով կամ սովորել եք այլ լեզուներ, նշե՛ք այդ լեզուները, որ տարիքից եք խոսում կամ սովորում և ինչքան երկար եք սովորել. _____ _____

Կրթության աստիճանը. _____	_____ _____

Հոր մասնագիտությունը և/կամ կատարած աշխատանքը/աշխատանքները.

_____

Մոր մասնագիտությունը և/կամ կատարած աշխատանքը/աշխատանքները.

_____

Ո՞ր երկրում/քաղաքում/գյուղում է ծնվել Ձեր հայրը: _____

Ո՞ր երկրում/քաղաքում/գյուղում է ծնվել Ձեր մայրը: _____

Դուք ամուսնացա՞ծ եք: _____

Եթե այո, ո՞ր երկրում/քաղաքում/գյուղում է ծնվել Ձեր ամուսինը/կինը:

_____

Ո՞ր լեզվով/լեզուներով/բարբառով/բարբառներով եք խոսում իրար հետ:

_____

Participant Information

Date: _____ Gender: _____ Year of birth: _____ Birthplace: _____ If you have lived in locations other than your birthplace, indicate when, where, and for how long: _____ _____ _____ Current residence: _____ Highest level of education: _____	Profession: _____ Current job: _____ Previous work (if different from above): _____ Language(s) spoken at home until adulthood: _____ If you speak other languages or have studied other languages, indicate from what age / for how long you have used or studied them: _____ _____ _____ _____

Father's profession and/or job(s): _____

Mother's profession and/or job(s): _____

In what country/city/village was your father born? _____

In what country/city/village was your mother born? _____

Are you married? _____

If yes, in what country/city/village was your spouse born? _____

If yes, what language(s)/dialect(s) do you speak with your spouse? _____

Appendix C: Consent Forms

Pre-experiment consent form:

Հետազոտության վերնագիրը. Հնչյունների ընկալումը և արտաբերումը Հայաստանի Հանրապետության Գավառ քաղաքում

HUM00267270

Հետազոտողը. Էմմա Պորտուգալ՝ Միչիգանի համալսարանի ասպիրանտ (էլեկտրոնային հասցեն՝ esantelm@umich.edu)

Գիտական ղեկավարները. Սավիտրի Նամպուտիրիպադ՝ Միչիգանի համալսարանի դասախոս (էլեկտրոնային հասցեն՝ savithry@umich.edu) և Յելենա Կրիվոկապիչ՝ Միչիգանի համալսարանի դասախոս (էլեկտրոնային հասցեն՝ jelenak@umich.edu)

Հետազոտող-օգնականը. Անահիտ Պետրոսյան (anahitp1991@yahoo.com)

Ներածություն

Այս հետազոտությունը անցկացնում են Միչիգանի համալսարանի աշխատակիցները, և ֆինանսավորվում է Միչիգանի համալսարանի Հայագիտության կենտրոնի և Րաքհամի անվան հետբուհական կրթության ֆակուլտետի կողմից: Հետազոտության նպատակն է ուսումնասիրել այն, թե ինչպես են Գավառի բնակիչները արտասանում և ընկալում հնչյունները: Հավաքված տվյալների վերլուծությունը օգտագործվելու է հետազոտողի՝ Էմմա Պորտուգալի ակադեմիական ծրագրի բոլոր պահանջները բավարարելու համար:

Ընթացակարգ

Հետազոտությունը բաղկացած է 3 մասերից: Առաջին մասի նպատակն է ուսումնասիրել տարբեր բառերի և հնչյունների Ձեր արտասանությունը: Երկրորդ և երրորդ մասերի նպատակն է ուսումնասիրել տարբեր բառերի և հնչյունների Ձեր ընկալումը:

Առաջին մասի ընթացքում Դուք կարդալու եք բառերի մի ցանկ, և Դուք ձայնագրվելու եք:

Երկրորդ մասի ընթացքում Դուք ականջակալով լսելու եք ձայնագրված բառերի զույգեր և ընտրելու եք, թե այդ բառերը Ձեր կարծիքով նո՞ւյն ձև են հնչում, թե՞ տարբեր: Ձեր պատասխանները տալու եք համակարգչի ստեղները սեղմելով:

Երրորդ մասի ընթացքում Դուք ականջակալով մեկ առ մեկ լսելու եք բառեր և ընտրելու եք, թե տվյալ բառի ձևը / արտասանությունը / հնչողությունը Ձեր կարծիքով հատկանշակա՞ն է Գավառի բարբառին, թե՞ ոչ: Ձեր պատասխանները տալու եք համակարգչի ստեղները սեղմելով:

Հետազոտության 3 մասերն անցնելուց հետո Դուք լրացնելու եք Ձեր մասին կարճ հարցաթերթիկ և համաձայնության հետ կապված մի լրացուցիչ թերթիկ: Ամբողջ գործընթացը տևելու է մոտ 1.5-2 ժամ:

Ռիսկեր/Անհարմարություններ

Այս հետազոտությանը մասնակցելը նվազագույն ռիսկ է ներկայացնում: Հնարավոր է, որ Դուք որոշ անհարմարություն զգաք հետազոտության ժամանակ, բայց այդ անհարմարությունը ավելին չի լինելու, քան ամենօրյա կյանքում գոյություն ունեցող անհարմարությունները: Հնարավոր է, որ Ձեր աչքերը հոգնեն համակարգչի էկրանին նայելուց, կամ կարող եք ավելի ընդհանուր հոգնածություն զգալ հետազոտության կրկնվող բնույթի պատճառով:

Փոքր ռիսկ կա, թե կխախտվի Ձեր տվյալների գաղտնիությունը: Նման դեպքում անմիջապես տեղեկացնելու ենք Միչիգանի համալսարանի էթիկայի հանձնաժողովի անձնակազմին, ու նրանք ձեռնարկելու են համապատասխան միջոցներ: Էթիկայի հանձնաժողովը համալսարանի այն բաժինն է, որը զբաղվում է հետազոտությունների էթիկայի հարցերով:

Օգուտ

Հետազոտության մասնակիցները անմիջական օգուտ չեն ստանում: Այնուամենայնիվ, հույս ունենք, որ Ձեր մասնակցության շնորհիվ հետազոտողները ավելի լավ են հասկանալու այն, թե ինչպես է արտասանվում և ընկալվում հայերենը Գավառում, և ավելի ընդհանուր բացահայտումներ են անելու մարդկային խոսքի արտաբերման և ընկալման հետ կապված:

Գաղտնիություն

Մասնակիցների տվյալները գաղտնի են պահվելու: Գիտական հրատարակություններում ու զեկույցներում հետազոտության արդյունքները ներկայացվելու են առանց այնպիսի անձնական տվյալների, որոնք կարող են օգտագործվել Ձեր ինքնությունը բացահայտելու համար: Հնարավոր է, որ հետազոտության ընթացքում հավաքված տվյալները հասանելի դարձնենք այնպիսի հետազոտողների, ովքեր հետազոտական թիմի անդամ չեն, բայց նրանց հասանելի չենք դարձնելու որևէ անձնական տվյալ, որը կարող է օգտագործվել Ձեր ինքնությունը բացահայտելու համար: Հնարավոր է նաև, որ Միչիգանի համալսարանի էթիկայի հանձնաժողովի ներկայացուցիչները և կամ հետազոտությունը ֆինանսավորող բաժնի/բաժինների ներկայացուցիչները խնդրեն տվյալները ուսումնասիրել՝ հետազոտության անվտանգությունը վերահսկելու նպատակով:

Համաձայնությունը վկայող թղթե փաստաթղթերը (որոնք բովանդակելու են Ձեր ստորագրությունը, անուն-ազգանունը, հասցեն և էլեկտրոնային հասցեն) և թղթե հարցաթերթիկները (որոնք բովանդակելու են որոշ տեղեկություններ Ձեր կյանքի և լեզվական փորձառությունների հետ կապված) պահպանվելու են կողպեքով փակված սենյակում: Այս նյութերի թվայնացված տարբերակները և մնացած բոլոր տվյալները պահպանվելու են գաղտնաբառով պաշտպանված համակարգիչներում, արտաքին կոշտ սկավառակներում և UM-Dropbox-ում: UM-Dropbox-ը անվտանգ, գաղտնաբառով պաշտպանված ծառայություն է, որը համալսարանի աշխատակիցները օգտագործում են թվային ֆայլեր պահպանելու համար:

Այս փաստաթղթի «Համաձայնություն» հատվածում Դուք կարող եք համաձայնություն տալ հետազոտությանը մասնակցելու, իսկ «Լրացուցիչ համաձայնություն» հատվածում, որը լրացնելու եք հետազոտությունից հետո, կարող եք համաձայնություն տալ Ձեր տվյալները հետազոտական թիմի կողմից տարբեր լրացուցիչ ձևերով օգտագործվելու համար: Եթե համաձայնություն չտաք այս լրացուցիչ օգտագործման ձևերի համար, միևնույն է, կարող եք մասնակցել հետազոտությանը:

Փոխհատուցում

Դուք 20 դոլար եք ստանալու հետազոտությանը մասնակցելու համար: Միչիգանի համալսարանի ֆինանսների բաժինը պահանջում է, որ մենք գրանցենք Ձեր անուն-ազգանունը, հասցեն և էլեկտրոնային հասցեն (եթե ունեք), որպեսզի համալսարանը կարողանա հետևել Ամերիկայի Միացյալ Նահանգների հարկային օրենսգրքի բոլոր կանոններին:

Մասնակցություն

Դուք կամավորական սկզբունքով եք մասնակցելու այս հետազոտությանը: Իրավունք ունեք դադարեցնելու Ձեր մասնակցությունը ցանկացած պահին: Եթե այլևս չեք ցանկանում մասնակցել հետազոտությանը, խնդրում ենք տեղեկացնել հետազոտական թիմի անդամներին: Ցանկացած պահին կարող եք խնդրել, որ ջնջենք Ձեր տվյալները: Հարցաթերթիկը լրացնելիս իրավունք ունեք ցանկացած հարցի չպատասխանելու:

Հարցեր

Հետազոտության հետ կապված հարցերի դեպքում կարող եք գրել Էմմա Պորտուգալին (հետազոտող) esantelm@umich.edu հասցեով, Սավիտրի Նամպուտիրիպադին (գիտական ղեկավար) savithry@umich.edu հասցեով, կամ Յելենա Կրիվոկապիչին (գիտական ղեկավար) jelenak@umich.edu հասցեով: Միչիգանի համալսարանի Էթիկայի հանձնաժողովի Առողջապահական և վարքագծային գիտությունների բաժինը գտել է, որ այս հետազոտությունը նվազագույն ռիսկ է ներկայացնում մասնակիցների համար: Հետևաբար, հետազոտությունը ազատված է Էթիկայի հանձնաժողովի շարունակական վերահսկողությունից:

Համաձայնություն

Ստորև ստորագրելիս ես հաստատում եմ, որ ես 18 տարեկան եմ կամ ավելի մեծ, որ ես կարդացել ու հասկացել եմ վերը ներկայացված տեղեկությունները և ցանկանում եմ կամավորական սկզբունքով մասնակցել այս հետազոտությանը:

Ստորև ստորագրելիս համաձայնություն եմ տալիս նաև հետևյալ կետերի համար.

- Իմ տվյալները հասանելի են լինելու հետազոտական թիմի անդամներին և կարող են հասանելի դարձվել ուրիշ հետազոտողների՝ առանց որևէ անձնական տվյալի, որը կարող է օգտագործվել իմ ինքնությունը բացահայտելու համար:
- Հետազոտության արդյունքները գիտական հրատարակություններում ու զեկույցներում ներկայացվելու են առանց որևէ անձնական տվյալի, որը կարող է օգտագործվել իմ ինքնությունը բացահայտելու համար:

Անուն-ազգանուն _____

Մասնակցի ստորագրությունը _____

Ամսաթիվ _____

Հասցե _____

Էլեկտրոնային հասցե (եթե ունեք) _____

Additional consent form:

Լրացուցիչ համաձայնություն

Խնդրում ենք ստորև գրել Ձեր անվան և ազգանվան սկզբնատառերը այն կետերի կողքին, որոնց համար ցանկանում եք համաձայնություն տալ: Այս կետերը վերաբերում են նրան, թե ինչպես կարող ենք օգտագործել Ձեր տվյալները հետագայում: Դուք կամավոր սկզբունքով եք համաձայնություն տալիս այս կետերին և ցանկացած կետի կարող եք համաձայնություն չտալ: Մենք օգտագործելու ենք Ձեր տվյալները ըստ Ձեր նշած նախընտրությունների: Հետազոտության արդյունքները միշտ ներկայացվելու են առանց որևէ տվյալի, որը կարող է օգտագործվել Ձեր ինքնությունը բացահայտելու համար:

1. Հետազոտողը կարող է թույլ տալ, որ ուրիշ գիտնականներ և ուսանողներ լսեն կարճ ձայնագրություններ Ձեր կարդացած բառերի ցանկից գիտաժողովների ժամանակ հնչող զեկույցների և գիտական դասախոսությունների ժամանակ: Հնարավոր է, որ նման զեկույցները տեսագրվեն և համացանցում տեղադրվեն:

   Համաձայն եմ _____

2. Հետազոտողը կարող է հետագա հետազոտությունների մասնակիցներին ներկայացնել առանձին ձայնագրված բառեր Ձեր կարդացած բառերի ցանկից:

   Համաձայն եմ _____

Եթե ցանկանում եք չեղարկել Ձեր համաձայնությունը վերը նշված կետերից որևէ մեկի համար, կարող եք ցանկացած պահին գրել հետազոտողին՝ Էմմա Պորտուգալին (esantelm@umich.edu):

Ստորև ստորագրելիս հաստատում եմ, որ կարդացել եմ վերը նշված կետերը և համաձայնություն եմ տալիս, որ իմ տվյալները կարող են օգտագործվել վերը նշված ձևերով:

Անուն-ազգանուն _____

Մասնակցի ստորագրությունը _____

Ամսաթիվ _____

Հետագա հետազոտությունների մասնակցելու հնարավորություն

Հնարավոր է, որ հետազոտական թիմը ցանկանա Ձեզ հրավիրել մասնակցելու հետագա հետազոտություններին, որոնք կարող են նման լինել այս հետազոտությանը կամ շատ

տարբերվել սրանից: Խնդրում ենք ստորև նշել, թե կցանկանայիք, որ հրավիրենք Ձեզ մասնակցելու հետագա հետազոտություններին:

_____ Այո՛, համաձայնություն եմ տալիս, որ հետազոտողն ինձ հրավիրի մասնակցելու հետագա հետազոտություններին:

_____ Ո՛չ, համաձայնություն չեմ տալիս, որ հետազոտողն ինձ հրավիրի մասնակցելու հետագա հետազոտություններին:

Անուն-ազգանուն _____

Մասնակցի ստորագրությունը _____ Ամսաթիվ _____

Pre-experiment consent form (English translation):

Title of study: Perception and production of speech sounds in Gavaṛ, Armenia

HUM00267270

Researcher: Emma Portugal, graduate student at University of Michigan (email: esantelm@umich.edu)

Faculty advisors: Savithry Namboodiripad, Assistant Professor at University of Michigan (email: savithry@umich.edu) and Jelena Krivokapić, Associate Professor at University of Michigan (email: jelenak@umich.edu)

Research assistant: Anahit Petrosyan (anahitp1991@yahoo.com)

Introduction

This is a study conducted by personnel from the University of Michigan with funding from the Rackham Graduate School and Center for Armenian Studies at the University of Michigan. The goal of the study is to investigate how residents of Gavaṛ, Armenia pronounce and perceive various speech sounds. Analysis of the data collected in this study will be used to fulfil the requirements of the academic program of the researcher (Emma Portugal) as a graduate student.

Procedures

This study will consist of an experiment with three parts. The goal of the first part is to learn about how you pronounce different words and sounds, while the goal of the second and third parts is to learn about how you perceive different words and sounds.

In the first part, you will be recorded (audio only) reading a list of words.

In the second part, you will hear pairs of recorded words through headphones and will choose whether you think the words sound the same or different, indicating your responses with the keys of a computer.

In the third part, you will hear words one at a time through headphones and will choose whether or not you think each word's form / pronunciation / the way it sounds is characteristic of Gavaṛ's dialect or not, indicating your responses with the keys of a computer.

After completing these three parts, you will fill out a short questionnaire about yourself, as well as an additional consent form. The entire process will last approximately 1.5 to 2 hours.

Risks/Discomforts

Participating in this study poses minimal risks. It is possible that you will feel some discomfort during the experiment, but this discomfort will not be greater than the discomforts that exist in everyday life. You may experience eye fatigue due to looking at a computer screen, or you may experience more general fatigue or boredom due to the repetitive nature of the tasks in the experiment.

There is a small risk that your data's confidentiality could be breached. In this case, we will immediately notify the staff of the Institutional Review Board (IRB) of University of Michigan, who will take appropriate measures. The IRB is the department of the university that deals with research ethics.

Benefits

There are no direct benefits for participants. However, thanks to your participation, we hope that researchers will gain a better understanding of how Armenian is spoken and perceived in Gavaṛ, as well as gaining more general insights into the perception and production of human speech.

Confidentiality

Participants' data will be kept confidential. In scientific publications and presentations, the results of the study will be presented without identifying information. Data collected in this project may be shared with researchers outside of the study team, but we will not share any information that could identify you. It is also possible that the IRB of University of Michigan and/or the department(s) funding the research may request to access the data in order to monitor the safety of the research.

Your paper consent forms (which will contain your signature, name, address, and email address) and paper questionnaire (which will contain some information about your life and your language experiences) will be kept in a locked room. Digitized versions of these materials, as well as all other data, will be stored in password-protected computers, external hard drives, and on UM-Dropbox. UM-Dropbox is a secure, password-protected service that university employees use to store digital files.

In the section of this form titled "Consent," you will be able to give your consent to participate in the experiment. In the "Additional Consent" section, which you will complete after the experiment, you will be able to give your consent for the research team to use your data in additional ways. If you do not give consent for these additional uses, you can still participate in the study.

Compensation

You will receive $20 for your participation in this study. The University of Michigan finance department will need your name, address, and email address for tax reporting purposes.

Participation

You will participate in this study voluntarily. You have the right to withdraw from the study at any time. If you no longer desire to participate in the study, please inform a member of the study team. You can ask us to delete your data at any time. While filling out the questionnaire, you have the right to skip any question.

Questions about the Research

If you have questions regarding this study, you may contact Emma Portugal (Principal Investigator) at esantelm@umich.edu, Savithry Namboodiripad (Faculty Advisor) at savithry@umich.edu, or Jelena Krivokapić (Faculty Advisor) at jelenak@umich.edu. The Health Sciences and Behavioral Sciences Institutional Review Board (IRB) of the University of Michigan has determined that this research poses no more than minimal risk and is exempt from ongoing IRB oversight.

Consent

By signing below, I confirm that I am 18 years or older, that I have read and understood the information presented above, and that I wish to voluntarily participate in this study.

By signing below, I also consent to the following:

- My data will be accessible to the research team, and may be shared with other researchers without any identifying information
- The results of the study will be presented in scientific presentations and publications without any identifying information

Print name _____

Signature of Participant _____ Date _____

Address _____

Email address (if you have one) _____

Additional consent form (English translation):

Additional Consent

Please initial below next to those points for which you wish to give your consent. These points pertain to the ways in which we will be able to use your data in the future. Giving your consent for these points is voluntary, and you can choose to not give consent to either of them. We will use your data in accordance with your indicated preferences. The results of the study will always be presented without identifying information.

1. The researcher can allow other researchers and students to hear short recordings of your speech (from the word list) during scientific lectures or conference presentations. It is possible that such presentations will be recorded and posted on the internet.

   I agree _____

2. The researcher can present recordings of your speech (individual words from the word list) to participants in future studies.

   I agree _____

If you wish to revoke your consent for either of the above items, you can write to the researcher, Emma Portugal (esantelm@umich.edu), at any time.

By signing below, I confirm that I have read the items listed above and give my consent for my data to be used in the ways that I have indicated.

Print name _____

Signature of Participant _____ Date _____

Opportunities to Participate in Future Research

It is possible that the research team may wish to invite you to participate in other studies in the future. These studies may be similar to the present study, or they may be very different. Please indicate below whether or not you would like to be invited to participate in future studies.

____ Yes, I give my consent for the researcher to invite me to participate in future studies.

____ No, I do not give my consent for the researcher to invite me to participate in future studies.

Print name _____

Signature of Participant _____ Date _____

Appendix D: Verbal Task Instructions

Armenian:

Հիմա սկսենք: Նախ՝ ես ձեզ համակարգչի էկրանով ցույց եմ տալու բառերի մի ցանկ: Դուք բարձրաձայն կարդալու եք այդ բառերը, և դուք ձայնագրվելու եք: Խնդրում ենք բառերը կարդալ այնպես, ինչպես ասում եք առօրյա կյանքում: Կարդալու ճիշտ կամ սխալ ձև չկա, քանի որ մենք ուղղակի ուզում ենք իմանալ, թե դուք ինչպես եք ասում բառերը: Բառերը կարդալուց հետո դուք անցնելու եք հետազոտության ևս երկու մաս, որոնց ընթացքում դուք լսելու եք ձայնագրված բառեր և պատասխաններ եք տալու՝ այդ բառերի հետ կապված ձեր տպավորությունների հիման վրա: Հետո ավելի մանրամասն կբացատրեմ ամեն մասի ընթացակարգը: Հետազոտության այդ մասերում չկան ճիշտ կամ սխալ պատասխաններ, քանի որ մենք ուղղակի ուզում ենք իմանալ, թե դուք ինչպես եք ընկալում բառերը և հնչյունները: Հնարավոր է, որ ձեզ երբեմն դժվար թվա պատասխան ընտրել, կամ ձեզ թվա, որ վստահ չեք, թե որ պատասխանն եք ուզում ընտրել: Նման դեպքերում խնդրում ենք պատասխան ընտրել ըստ ձեր առաջին տպավորության: Դուք չեք ունենալու հնարավորություն հետ գնալու, բառերը նորից լսելու կամ փոխելու ձեր պատասխանները, այնպես որ ուշադի՛ր եղեք:

Հիմա դուք անցնելու եք հետազոտության առաջին մասը, որի ընթացքում դուք լսելու եք ձայնագրված բառեր և պատասխաններ եք տալու՝ ըստ ձեր տպավորությունների՝ այդ բառերի հետ կապված: Ձայնագրված բառերը մշակվել են համակարգչային ծրագրի օգնությամբ: Դուք լսելու եք որոշ բառերի արտասանության երկու ձևեր և փորձելու եք հասկանալ, թե արդյոք այդ բառերի հնչողությունները նո՞ւյնն են, թե՞ տարբեր են: Եթե կարծում եք, որ դրանք նույնն են, սեղմե՛ք 1: Եթե կարծում եք, որ դրանք տարբեր են, սեղմե՛ք 0: Հիշե՛ք, որ ճիշտ կամ սխալ պատասխաններ չկան, քանի որ մենք ուղղակի ուզում ենք իմանալ, թե դուք ինչպես եք ընկալում բառերը և հնչյունները: Հնարավոր է, որ ձեզ երբեմն դժվար թվա պատասխան ընտրել, կամ ձեզ թվա, որ վստահ չեք, թե որ պատասխանն եք ուզում ընտրել: Նման դեպքերում խնդրում ենք պատասխան ընտրել ըստ ձեր առաջին տպավորության: Դուք չեք ունենալու հնարավորություն հետ գնալու, բառերը նորից լսելու կամ փոխելու ձեր պատասխանները, այնպես որ ուշադի՛ր եղեք:

[**Կարդացե՛ք այս մասը, երբ հետազոտության ծրագիրը արդեն բացված է, և վարժանքային մասի հրահանգները էկրանին են**] Նախ՝ դուք անցնելու եք վարժանքային կարճ հատված, որպեսզի հետազոտության ընթացքին վարժվեք: Եթե գտնեք, որ ձայնը չափազանց բարձր կամ ցածր է, կարող եք այն կարգավորել այս հատվածի ընթացքում [**ցո՛ւյց տվեք ձայնի կարգավորման ձևը**]: Երբ նորից տեսնեք էկրանին հրահանգները, վարժանքային հատվածը ավարտված կլինի, և հետազոտության հիմնական հատվածը կսկսվի:

[**Կարդադե՛ք այս մասը ընդմիջման ժամանակ**] Հիմա դուք անցնելու եք հետազոտության երկրորդ մասը, որի ընթացքում դուք լսելու եք ձայնագրված բառեր և պատասխաններ եք տալու՝ ըստ ձեր տպավորությունների՝ այդ բառերի հետ կապված: Ձայնագրված բառերը մշակվել են համակարգչային ծրագի օգնությամբ: Դուք մեկ առ մեկ լսելու եք որոշ բառեր և որոշելու եք, թե այդ բառերի ձևը / արտասանությունը / հնչողությունը Գավառի բարբառին հատկանշակա՞ն է, թե՞ ոչ: Եթե կարծում եք, որ հատկանշական է Գավառի բարբառին, սեղմե՛ք 1: Եթե կարծում եք, որ հատկանշական չէ Գավառի բարբառին (այսինքն՝ այդպես կարող են ասել ոչ միայն գավառցիները, այլև Հայաստանի այլ բնակիչները), սեղմե՛ք 0: Բոլոր բառերը նույն մարդն է արտասանելու: Դրա համար որոշումը պիտի կայացնեք ոչ թե մարդու ձայնի հիման վրա, այլ բառի ձևի / արտասանության / հնչողության հիման վրա: Հիշե՛ք, որ ճիշտ կամ սխալ պատասխաններ չկան, քանի որ մենք ուղղակի ուզում ենք իմանալ, թե դուք ինչպես եք ընկալում բառերը և հնչյունները: Հնարավոր է, որ ձեզ երբեմն դժվար թվա պատասխան ընտրել , կամ ձեզ թվա, որ վստահ չեք, թե որ պատասխանն եք ուզում ընտրել: Նման դեպքերում խնդրում ենք պատասխան ընտրել ըստ ձեր առաջին տպավորության: Դուք չեք ունենալու հնարավորություն հետ գնալու, բառերը նորից լսելու կամ փոխելու ձեր պատասխանները, այնպես որ ուշադի՛ր եղեք:

[**Կարդացե՛ք այս մասը, երբ վարժանքային մասի հրահանգները էկրանին են**] Նախ՝ դուք անցնելու եք վարժանքային կարճ հատված, որպեսզի հետազոտության ընթացքին վարժվեք: Երբ նորից տեսնեք էկրանին հրահանգները, վարժանքային հատվածը ավարտված կլինի, և հետազոտության հիմնական հատվածը կսկսվի:

English translation:

Now let’s begin. First, I will show you a list of words on the computer screen. You will read these words aloud and will be recorded. Please read the words as you say them in everyday life. There is no correct or incorrect way of reading, since we simply want to understand how you say the words. After reading the words you will complete two more parts of the study, during which you will hear recorded words and will give answers according to your impressions about those words. Later on, I will explain the procedure for each part in more detail. In those parts of the study, there are no correct or incorrect answers, since we simply want to understand how you perceive words and sounds. It is possible that you will sometimes find it difficult to choose an answer, or that you will feel unsure about which answer you want to choose. In such

cases, please choose an answer based on your first impression. You will not be able to go back, hear the words again, or change your answers, so listen carefully.

Now you will complete the first part of the study during which you will hear recorded words and give answers according to your impressions about those words. The recorded words have been modified using a computer program. You will hear two forms of the pronunciation of some words and will try to decide whether those words sound the same or different. If you think that they are the same, press 1. If you think that they are different, press 0. Remember that there are no correct or incorrect answers, since we simply want to understand how you perceive words and sounds. It is possible that you will sometimes find it difficult to choose an answer, or that you will feel unsure about which answer you want to choose. In such cases, please choose an answer based on your first impression. You will not be able to go back, hear the words again, or change your answers, so listen carefully.

[Read this part when the experiment software is already open, and the instructions for the training portion are on the screen] First, you will complete a short training portion, in order to get used to the study procedure. If you find that the sound is too loud or too quiet, you can adjust it during this part **[show how to adjust the volume]**. When you see the instructions on the screen again, the training portion will be over, and the main portion of the study will begin.

[Read this part during the break] Now you will complete the second part of the study during which you will hear recorded words and give answers according to your impressions about those words. The recorded words have been modified using a computer program. You will hear words one at a time and decide if those words' form / pronunciation / the way they sound is characteristic of Gavaṛ's dialect or not. If you think that it is characteristic of Gavaṛ's dialect, press 1. If you think that it is not characteristic of Gavaṛ's dialect (that is, not only people from Gavaṛ, but also other residents of Armenia can speak that way), press 0. The same person will pronounce all of the words. So, you should make your decision not based on the person's voice, but rather based on the word's form / pronunciation / the way it sounds. Remember that there are no correct or incorrect answers, since we simply want to understand how you perceive words and sounds. It is possible that you will sometimes find it difficult to choose an answer, or that you will feel unsure about which answer you want to choose. In such cases, please choose an answer based on your first impression. You will not be able to go back, hear the words again, or change your answers, so listen carefully.

[Read this part when the experiment software is already open, and the instructions for the training portion are on the screen] First, you will complete a short training portion, in order to get used to the study procedure. When you see the instructions on the screen again, the training portion will be over, and the main portion of the study will begin.

Appendix E: Individual Responses to Dialect Categorization Task

In the figures below, purple represents "Gavaṛ" responses (1), while yellow represents "not Gavaṛ" responses (0).

Appendix Figure E.1: Individual responses for *pat*



Appendix Figure E.2: Individual responses for *mas*



Appendix Figure E.3: Individual responses for *mard*



Appendix Figure E.4: Individual responses for *pos*



Appendix Figure E.5: Individual responses for *tsots*

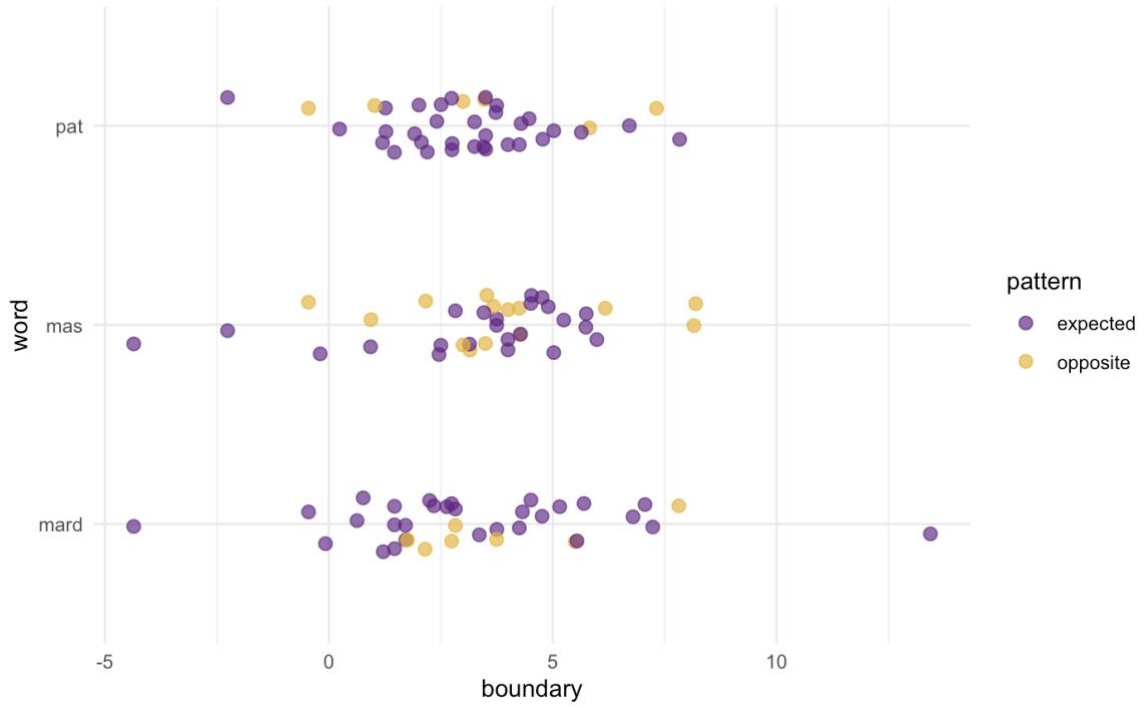


Appendix Figure E.6: Individual responses for *xoz*

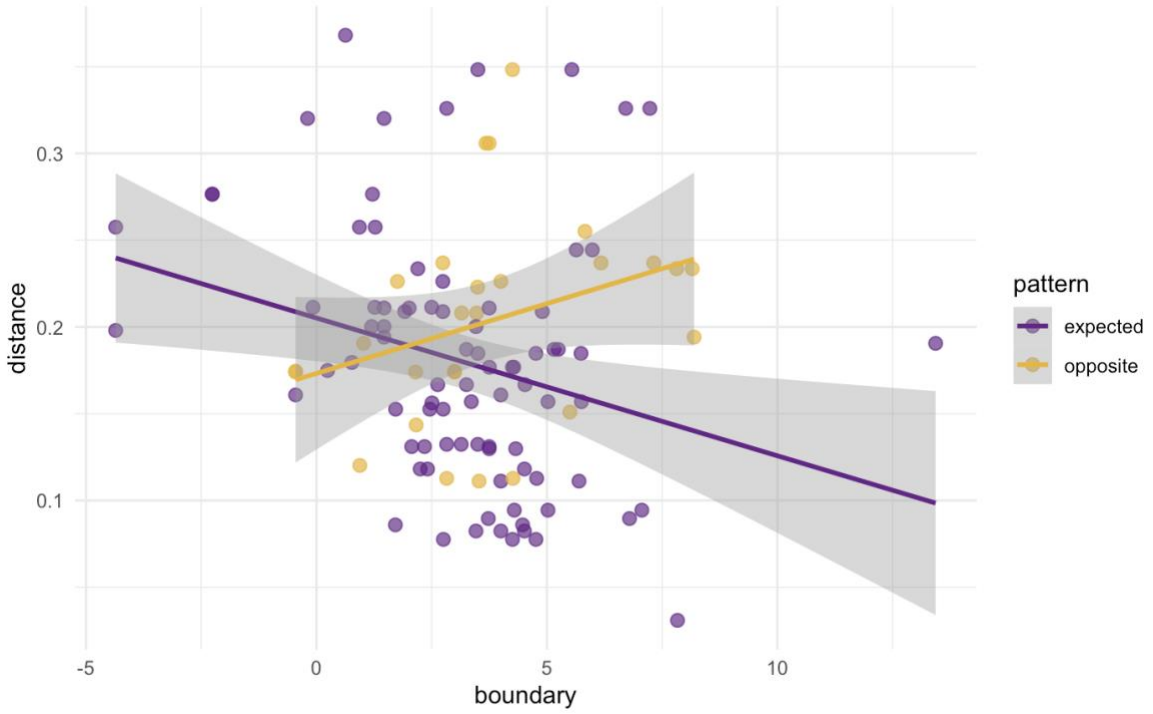


Appendix F: Additional Figures Related to BOUNDARY LOCATION

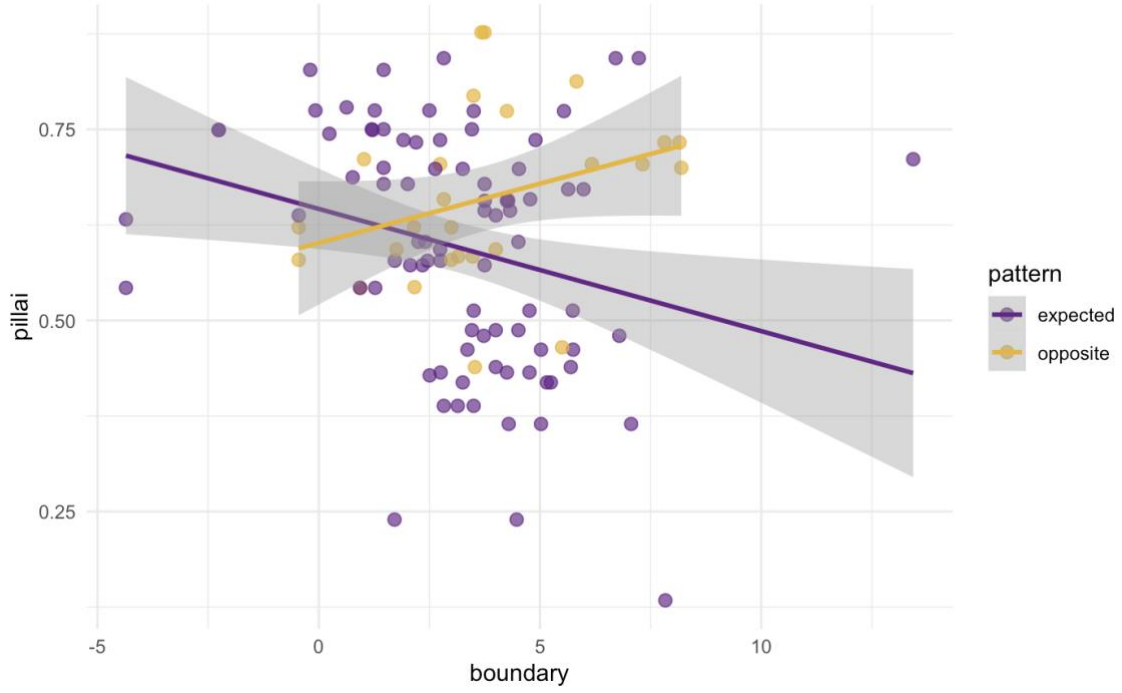
Appendix Figure F.1: Dialect categorization boundary locations by word for the [a]/[ɒ]$_A$ vowel pair. The color coding indicates participants who had response patterns with expected directionality (i.e., categorizing the dialect-like end of the continuum as dialectal and the SEA-like end of the continuum as non-dialectal) vs. those whose response pattern had the opposite directionality. Each dot represents one participant.



Appendix Figure F.2: Dialect categorization boundary locations for the [a]/[ɒ]$_A$ vowel pair (x-axis) in relation to ED-median (from non-Bark-converted data) from participants' productions of the same vowel pair (y-axis) with overlaid regression lines. The color coding indicates participants who had response patterns with expected directionality (i.e., categorizing the dialect-like end of the continuum as dialectal and the SEA-like end of the continuum as non-dialectal) vs. those whose response pattern had the opposite directionality. Each dot represents one participant's boundary location for one word.



Appendix Figure F.3: Dialect categorization boundary locations for the [a]/[ɒ]$_{A}$ vowel pair (x-axis) in relation to Pillai scores from participants' productions of the same vowel pair (y-axis) with overlaid regression lines. The color coding indicates participants who had response patterns with expected directionality (i.e., categorizing the dialect-like end of the continuum as dialectal and the SEA-like end of the continuum as non-dialectal) vs. those whose response pattern had the opposite directionality. Each dot represents one participant's boundary location for one word.



Bibliography

Abeghyan, Manuk. (1971). *Erger* [Works] (vol. 5). Yerevan: Haykakan SSH GA Hratarakch^{h}ut^{h}yun [Publishing House of the Academy of Sciences of the Armenian SSR].

Abrahamian, Levon H. (1998). Mother tongue: Linguistic nationalism and the cult of translation in postcommunist Armenia. *Berkeley Program in Soviet and Post-Soviet Studies Working Paper Series*. https://escholarship.org/uc/iseees_bps_rw/search?sort=a-author&rows=10&start=0.

Abrahamian, Levon H. (2006). *Armenian identity in a changing world.* Costa Mesa California: Mazda Publishers.

Acharyan, Hrach^{h}ya. (1911). *Hay barbaṛagitut^{h}yun* [Armenian dialectology]. Moscow: Lazaryan Chemaran Arevelyan Lezvats [Lazaryan Seminary of Eastern Languages].

Adank, Patti, Roel Smits & Roeland van Hout. (2004). A comparison of vowel normalization procedures for language variation research. *Journal of the Acoustical Society of America, 116*, 3099–3107. https://doi.org/10.1121/1.1795335.

Ainsworth, W.A. & K.K. Paliwal. (1984). Correlation between the production and perception of the English glides /w, r, l, j/. *Journal of Phonetics* 12. 237-243. https://doi.org/10.1016/S0095-4470(19)30880-0.

Allen, W. Sidney. (1950). Notes on the phonetics of an Eastern Armenian speaker. *Transactions of the Philological Society* 49 (1). 180-206. https://doi.org/10.1111/j.1467-968X.1950.tb00241.x.

Amengual, Mark & Pilar Chamorro. (2015). The effects of language dominance in the perception and production of the Galician mid vowel contrasts. *Phonetica* 72. 207-236. https://doi.org/10.1159/000439406.

Arnold, Lacey R. (2015). Multiple mergers: Production and perception of three pre-/l/ mergers in Youngstown, Ohio. *University of Pennsylvania Working Papers in Linguistics* 21(2). 1-10.

Auer, Peter. (2005). Europe's sociolinguistic unity, or: A typology of European dialect/standard constellations. In Nicole Delbecque, Johan van der Auwera & Dirk Geeraerts (eds.), *Perspectives on Variation: Sociolinguistic, Historical, Comparative*, 7-42. Berlin, New York: De Gruyter Mouton. https://doi.org/10.1515/9783110909579.7.

Austen, Martha. (2020). Production and perception of the PIN-PEN merger. *Journal of Linguistic Geography* 8(2). 115-126. https://doi.org/10.1017/jlg.2020.9.

Babel, Anna M. (2016). Preface. In Anna M. Babel (Ed.), *Awareness and control in sociolinguistic research* (pp. xix-xxii). Cambridge University Press.

Babel, Molly. (2009). Selective vowel imitation in spontaneous phonetic accommodation. *UC Berkeley Phonology Lab Annual Report* 5(5). 163-194. https://escholarship.org/uc/item/6c95w485.

Babel, Molly. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics* 40. 177-189. https://doi.org/10.1016/j.wocn.2011.09.001.

Babel, Molly, Michael McAuliffe & Graham Haber. (2013). Can mergers-in-progress be unmerged in speech accommodation? *Frontiers in Psychology* 4. 653. https://doi.org/10.3389/fpsyg.2013.00653.

Baese-Berk, Melissa M., Efthymia C. Kapnoula & Arthur G. Samuel. (2025). The relationship of speech perception and speech production: It's complicated. *Psychonomic Bulletin & Review* 32(1). 226-242. https://doi.org/10.3758/s13423-024-02561-w.

Baghramyan, Ṛuben Harut^{h}yuni. (1972). *Hayeren barbaṛner: Sevana lchi avazani khosvatsk^{h}nerě* [Armenian dialects: Dialects of the Lake Sevan basin]. Yerevan: HSSH bardzraguyn ev mijnakarg masnagitakan krt^{h}ut^{h}yan ministrut^{h}yun, Erevani Kh. Abovyani anvan haykakan petakan mankavarzhakan institut [ASSR Ministry of Higher and Secondary Professional Education, Armenian State Pedagogical Institute of Yerevan named after Kh. Abovyan].

Bailey, Peter J. & Mark P. Haggard. (1973). Perception and production: Some correlations on voicing of an initial stop. *Language and speech* 16(3). 189-195. https://doi.org/10.1177/002383097301600301.

Baranowski, Maciej. (2013). On the role of social factors in the loss of phonemic distinctions. *English Language and Linguistics* 17(2). 271-295. https://doi.org/10.1017/S1360674313000038.

Baranowski, Maciej. (2022). Part of town as an independent factor: The NORTH-FORCE merger in Manchester. *Language Variation and Change* 34. 239-269. https://doi.org/10.1017/S095439452200014X.

Barreda, Santiago. (2021). Fast Track: Fast (nearly) automatic formant-tracking using Praat. *Linguistics Vanguard* 7(1). 20200051. https://doi.org/10.1515/lingvan-2020-0051.

Bates Douglas, Martin Mächler, Ben Bolker & Steve Walker. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1-48. https://doi.org/10.18637/jss.v067.i01.

Beaton, Mary Elizabeth. (2016). Revisiting Incomplete Neutralization: The Case of Puerto Rican Spanish. *University of Pennsylvania Working Papers in Linguistics* 22(1). 31-40.

Becker, Kara. (2017). The sociolinguistic interview. In Christine Mallinson, Becky Childs & Gerard Van Herk (eds.), *Data collection in sociolinguistics: Methods and applications*, 2nd edn., 99-107. London & New York: Routledge.

Beddor, Patrice Speeter. (2015). The relation between language users' perception and production repertoires. *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow, UK (Keynote talk). https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS1041.pdf.

Beddor, Patrice Speeter, Andries W. Coetzee, Will Styler, Kevin B. McGowan, & Julie E. Boland. (2018). The time course of individuals' perception of coarticulatory information is linked to their production: Implications for sound change. *Language* 94(4). 931-968. https://doi.org/10.1353/lan.2018.0051.

Beddor, Patrice Speeter, Andries W. Coetzee, Ian Calloway, Stephen Tobin & Ruaridh Purse. (2024). The relation between perceptual retuning and articulatory restructuring: Individual differences in accommodating a novel phonetic variant. *Journal of Phonetics* 107. 101352. https://doi.org/10.1016/j.wocn.2024.101352.

Beddor, Patrice Speeter, James D. Harnsberger, & Stephanie Lindemann. (2002). Language-specific patterns of vowel-to-vowel coarticulation: acoustic structures and their

perceptual correlates. *Journal of Phonetics* 30. 591-627. https://doi.org/10.1006/jpho.2002.0177.
Beddor, Patrice Speeter, Kevin B. McGowan, Julie E. Boland, Andries W. Coetzee, & Anthony Brasher (2013). The time course of perception of coarticulation. *Journal of the Acoustical Society of America* 133. 2350-2366. https://doi.org/10.1121/1.4794366.
Beddor, Patrice Speeter & Rena Arens Krakow. (1999). Perception of coarticulatory nasalization by speakers of English and Thai: Evidence for partial compensation. *Journal of the Acoustical Society of America* 106. 2868-2887. https://doi.org/10.1121/1.428111.
Bell, Alan. (1984). Language Style as Audience Design. *Language in Society* 13(2). 145-204. http://www.jstor.org/stable/4167516.
Bell-Berti, Fredericka, Lawrence J. Raphael, David B. Pisoni, & James R. Sawusch. (1979). Some relationships between speech production and perception. *Phonetica* 36. 373-383. https://doi.org/10.1159/000259974.
Benus, Stefan & Adamantios Gafos. (2007). Articulatory characteristics of Hungarian 'transparent' vowels. *Journal of Phonetics* 35. 271-300. https://doi.org/10.1016/j.wocn.2006.11.002.
Berry, Grant M. & Mirjam Ernestus. (2018). Phonetic alignment in English as a lingua franca: Coming together while splitting apart. *Second Language Research* 34(3). 343-370. https://doi.org/10.1177/0267658317737348.
Bochynska, Agata, Liam Keeble, Caitlin Halfacre, Joseph V. Casillas, Irys-Amélie Champagne, Kaidi Chen, Melanie Röthlisberger, Erin M. Buchanan & Timo B. Roettger (2023). Reproducible research practices and transparency across linguistics. *Glossa Psycholinguistics* 2(1):18. 1-36. https://doi.org/10.5070/G6011239.
Boersma, Paul. (1998). *Functional Phonology*. The Hague: Holland Academic Graphics.
Boersma, Paul & David Weenink. (2021). Praat: doing phonetics by computer [Computer program]. Version 6.1.52. http://www.praat.org/.
Boomershine, Amanda, Kathleen Currie Hall, Elizabeth Hume, & Keith Johnson. (2008). The impact of allophony vs. contrast on speech perception. In Peter Avery, B. Elan Dresher & Keren Rice (eds.), *Contrast in phonology: Theory, perception, acquisition*, 145-171. De Gruyter Mouton. https://doi.org/10.1515/9783110208603.2.145.
Bowie, David. (2000). Perception and production in a series of related mergers. *Linguistic Association of Canada and the United States Forum* 27. 297-305. https://go.gale.com/ps/i.do?id=GALE%7CA307270800&sid=googleScholar&v=2.1&it=r&linkaccess=abs&issn=0195377X&p=LitRC&sw=w&userGroupName=umuser&aty=i.
Bowie, David (2021). Individual and group trajectories across adulthood in a sample of Utah English speakers. In Karen V. Beaman & Isabelle Buchstaller (eds.), *Language variation and language change across the lifespan*, 101-118. New York: Routledge.
Britain, David & Sarah Grossenbacher. (2021). Counterurbanisation, dialect contact and the levelling of non-salient traditional dialect variants. In Arne Ziegler, Stefanie Edler & Georg Oberdorfer (eds.), *Urban matters: Current approaches in variationist sociolinguistics*, 89-118. John Benjamins Publishing Company.
Butcher, Kerri-Ann. (2022). *The MOAN-MOWN and MOAN-GOOSE mergers in Lowestoft English: Perception and production* [Unpublished doctoral dissertation]. University of Cambridge.
Campbell-Kibler, Kathryn. (2012). The Implicit Association Test and sociolinguistic meaning. *Lingua* 122. 753-763. https://doi.org/10.1016/j.lingua.2012.01.002.

Campbell-Kibler, Kathryn. (2016). Towards a Cognitively Realistic Model of Meaningful Sociolinguistic Variation. In Anna M. Babel (ed.), *Awareness and control in sociolinguistic research*,123-151. Cambridge University Press.
Carignan, Christopher. (2018). Using ultrasound and nasalance to separate oral and nasal contributions to formant frequencies of nasalized vowels. *The Journal of the Acoustical Society of America 143*(5). 2588. https://doi.org/10.1121/1.5034760.
Casasanto, Laura Staum. (2010). What do listeners know about sociolinguistic variation? *University of Pennsylvania Working Papers in Linguistics* 15(2). 40-49.
Chambers, J.K. & Peter Trudgill. (1998). *Dialectology,* 2nd edn. Cambridge: Cambridge University Press.
Chevalier, Alida. (2016). *Globalisation versus internal development: The reverse short front vowel shift in South African English* [Unpublished doctoral dissertation]. University of Cape Town.
Clopper, Cynthia. G. (2024). Dynamic acoustic vowel distances within and across dialects. *Journal of the Acoustical Society of America* 156. 2497-2507. https://doi.org/10.1121/10.0032385.
Coetzee, Andries W., Patrice Speeter Beddor, Will Styler, Stephen Tobin, Ian Bekker & Daan Wissing. (2022). Producing and perceiving socially structured coarticulation: Coarticulatory nasalization in Afrikaans. *Laboratory Phonology* 13(1). 13. https://doi.org/10.16995/labphon.6450.
Cole, Amanda & Patrycja Strycharczuk (2019). 'The price-mouth crossover in the "Cockney diaspora." In *Proceedings of the 19th International Congress of Phonetic Sciences*, 602-606.
Corretge, Ramon (2012-2024). Praat Vocal Toolkit. https://www.praatvocaltoolkit.com.
Costa, Paul & Ignatius G. Mattingly. (1981). Production and perception of phonetic contrast during phonetic change. *Journal of the Acoustical Society of America* 69. S67. https://doi.org/10.1121/1.386167.
Cowe, S. Peter. (1992). Amēn teł hay kay: Armenian as a pluricentric language. In Michael Clyne (ed.), *Pluricentric Languages: Differing Norms in Different Nations*, 325-345. Berlin: Mouton de Gruyter.
Cox, Felicity, Joshua Penney & Sallyanne Palethorpe. (2024). Australian English monophthong change across 50 years: Static versus dynamic measures. *Languages* 9(3). 99. https://doi.org/10.3390/languages9030099.
Diehl, Randy L, Björn Lindblom, Kathryn A. Hoemeke & Richard P. Fahey. (1996). On explaining certain male-female differences in the phonetic realization of vowel categories. *Journal of Phonetics* 24. 187-208. https://doi.org/10.1006/jpho.1996.0011.
Di Paolo, Marianna. (1988). Pronunciation and categorization in sound change. In Kathleen Ferrara, Becky Brown, Keith Walters, & John Baugh (eds.), *Linguistic change and contact: proceedings of the Sixteenth Annual Conference on New Ways of Analyzing Variation*, 84-92.
Di Paolo, Marianna. (1992). Hypercorrection in response to the apparent merger of (ɔ) and (ɑ) in Utah English. *Language & Communication* 12(3/4). 267-292. https://doi.org/10.1016/0271-5309(92)90017-4.
Di Paolo, Marianna & Alice Faber. (1990). Phonation differences and the phonetic content of the tense-lax contrast in Utah English. *Language Variation and Change* 2. 155-204. https://doi.org/10.1017/S0954394500000326.

D'Onofrio, Annette. (2018). Controlled and automatic perceptions of a sociolinguistic marker. *Language Variation and Change* 30. 261-285. https://doi.org/10.1017/S095439451800008X.
D'Onofrio, Annette. (2021). Age-based perceptions of a reversing regional sound change. *Journal of Phonetics* 86. https://doi.org/10.1016/j.wocn.2021.101038.
D'Onofrio, Annette, Teresa Pratt & Janneke Van Hofwegen. (2019). Compression in the California Vowel Shift: Tracking generational sound change in California's Central Valley. *Language Variation and Change* 31. 193-217. https://doi.org/10.1017/S0954394519000085.
Dossey, Ellen, Cynthia G. Clopper & Laura Wagner. (2020). The development of sociolinguistic competence across the lifespan: Three domains of regional dialect perception. *Language Learning and Development* 16(4). 330-350. https://doi.org/10.1080/15475441.2020.1784736.
Drager, Katie. (2005). From bad to bed: the relationship between perceived age and vowel perception in New Zealand English. *Te Reo* 48. 55-68.
Dum-Tragut, Jasmine. (2009). *Armenian: Modern Eastern Armenian.* John Benjamins Publishing Company.
Dum-Tragut, Jasmine. (2013). Striving for linguistic independence? The Armenian language in Post-Soviet Armenia: Language policy and language planning. In Sarah Smyth & Conny Opitz (eds.), *Negotiating linguistic, cultural and social identities in the post-Soviet world*, 279-300. Bern, Switzerland: Peter Lang.
Eckert, Penelope. (2012). Three Waves of Variation Study: The Emergence of Meaning in the Study of Sociolinguistic Variation. *Annual Review of Anthropology* 41. 87-100. https://doi.org/10.1146/annurev-anthro-092611-145828.
Eckert, Penelope & William Labov. (2017). Phonetics, phonology and social meaning. *Journal of Sociolinguistics* 21(4). 467-496. https://doi.org/10.1111/josl.12244.
Emmert-Streib, Frank, Salissou Moutari & Matthias Dehmer. (2023). *Elements of data science, machine learning, and artificial intelligence using R*. Cham, Switzerland: Springer.
Etemady, Dara L.J. & Timo B. Roettger. (2025). Statistical reporting inconsistencies in experimental linguistics. *Glossa Psycholinguistics* 4(1):21. 1-18. https://doi.org/10.5070/G6011.47896.
Evanini, Keelan. (2009). *The permeability of dialect boundaries: a case study of the region surrounding Erie, Pennsylvania* [Unpublished doctoral dissertation]. University of Pennsylvania.
Evans, Bronwen G. & Paul Iverson. (2007). Plasticity in vowel perception and production: A study of accent change in young adults. *Journal of the Acoustical Society of America* 121. 3814-3826. https://doi.org/10.1121/1.2722209.
Flanagan, James L. (1955). A difference limen for vowel formant frequency. *Journal of the Acoustical Society of America* 27(3). https://doi.org/10.1121/1.1907979.
Ferguson, Charles. (1959). Diglossia. *Word* 15. 325-340.
Forsyth, David. (2018). *Probability and statistics for computer science*. Springer.
Fowler, Carol A. (2006). Compensation for coarticulation reflects gesture perception, not spectral contrast. *Perception & Psychophysics* 68(2). 161-177. https://doi.org/10.3758/BF03193666.
Fox, John & Sanford Weisberg. (2019). *An R companion to applied regression*, 3rd edn. Thousand Oaks, CA: Sage. https://www.john-fox.ca/Companion/.

Fox, Robert Allen. (1982). Individual variation in the perception of vowels: Implications for a perception-production link. *Phonetica* 39. 1-22. https://doi.org/10.1159/000261647.
Freeman, Valerie & Paul De Decker. (2021). Remote sociophonetic data collection: Vowels and nasalization from self-recordings on personal devices. *Language and Linguistics Compass, 15*(7). https://doi.org/10.1111/lnc3.12435.
Freeman, Valerie. (2023). Production and perception of prevelar merger: Two-dimensional comparisons using Pillai scores and confusion matrices. *Journal of Phonetics* 97. https://doi.org/10.1016/j.wocn.2023.101213.
Fridland, Valerie & Tyler Kendall. (2012). Exploring the relationship between production and perception in the mid front vowels of U.S. English. *Lingua* 122. 779-793. https://doi.org/10.1016/j.lingua.2011.12.007.
Frieda, Elaina M., Amanda C. Walley, James E. Fledge, Michael E. Sloane. (2000). Adults' perception and production of the English vowel /i/. *Journal of Speech, Language, and Hearing Research* 43. 129-143. https://doi.org/10.1044/jslhr.4301.129.
Galstyan, Anahit (ed.). (2010). *Hayereni barbaṛagitakan atlas (Usumnasirut^{h}yunner yev nyut^{h}er): Prak 3* [Dialectological atlas of Armenian (Studies and materials): Part 3]. Yerevan: Nairi.
Garibova, Jala. (2020). Russian in Azerbaijan: Changing practices and emerging paradigms. In Arto Mustajoki, Ekaterina Protassova & Maria Yelenevskaya (eds.), *The soft power of the Russian language: Pluricentricity, politics, and policies*, 55-70. London: Routledge.
Garrett, Andrew. (1998). Adjarian's law, the glottalic theory, and the position of Armenian. In *Proceedings of the 24th annual meeting of the Berkeley linguistics society: Special session on Indo-European subgrouping and internal relations*, 12-23. Berkeley, CA. https://doi.org/10.3765/bls.v24i2.1250.
Ghazaryan, Lilit. (2020). *"Speak beautifully" – Language policies and practices in public kindergartens in Armenia.* Los Angeles: University of California, Los Angeles MA thesis. https://escholarship.org/uc/item/76h0k1bq.
Graff, David, William Labov, & Wendell A. Harris. (1986). Testing listeners' reactions to phonological markers of ethnic identity: A new method for sociolinguistic research. In David Sankoff (ed.), *Diversity and Diachrony*, 45-58. John Benjamins Publishing Company.
Gordon, Matthew J. (2013). Investigating chain shifts and mergers. In J.K. Chambers & Natalie Schilling (eds.), *The handbook of language variation and change* (2nd ed., pp. 203-219). John Wiley & Sons.
Grosvald, Michael & David Corina. (2012). The production and perception of sub-phonemic vowel contrasts and the role of the listener in sound change. In Maria-Josep Solé & Daniel Recasens (eds.), *The initiation of sound change: Perception, production, and social factors*, 77-100. John Benjamins Publishing Company.
Haeri, Niloofar. (1996). *The sociolinguistic market of Cairo: Gender, class and education*. London: Kegan Paul International.
Haghverdi, Vartan. (2016). *Armenian schwa: A phonetic and phonological analysis* [Unpublished MA thesis]. New Brunswick, NJ: Rutgers University. https://rucore.libraries.rutgers.edu/rutgers-lib/51304/.
Hall-Lew, Lauren. (2010). Improved representation of variance in measures of vowel merger. *Proceedings of Meetings on Acoustics* 9. 060002. https://doi.org/10.1121/1.3460625.

Harrington, Jonathan, Felicitas Kleber, Ulrich Reubold, Florian Schiel & Mary Stevens. (2019). The phonetic basis of the origin and spread of sound change. In William F. Katz & Peter F. Assmann (eds.), *The Routledge handbook of phonetics*, 401-426. New York: Routledge.

Harris, Kelly C., John H. Mills, Ning-Ji He & Judy R. Dubno. (2008). Age-related differences in sensitivity to small changes in frequency assessed with cortical evoked potentials. *Hear Res* 243(1-2): 47-56. https://doi.org/10.1016/j.heares.2008.05.005.

Hay, Jennifer, Katie Drager, & Brynmor Thomas. (2013). Using nonsense words to investigate vowel merger. *English Language and Linguistics* 17(2). 241-269. https://doi.org/10.1017/S1360674313000026.

Hay, Jennifer, Paul Warren, & Katie Drager. (2006). Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics* 34. 458-484. https://doi.org/10.1016/j.wocn.2005.10.001.

Heinze, Georg, Meinhard Ploner, Lena Jiricka & Gregor Steiner. (2025). *logistf: Firth's bias-reduced logistic regression*. R package version 1.26.1. https://github.com/georgheinze/logistf.

Helms, Annie. (2023). Sociophonetic analysis of mid front vowel production in Barcelona. In Barbara E. Bullock, Cinzia Russi & Almeida Jacqueline Toribio (eds.), *A half century of Romance linguistics: Selected proceedings of the 50th Linguistic Symposium on Romance Languages,* 249-268. Berlin: Language Science Press. https://doi.org/10.5281/zenodo.7525114.

Herold, Ruth. (1990). *Mechanisms of merger: The implementation and distribution of the low back merger in eastern Pennsylvania* [Unpublished doctoral dissertation]. University of Pennsylvania.

Herold, Ruth. (1997). Solving the actuation problem: Merger and immigration in eastern Pennsylvania. *Language Variation and Change* 9. 165-189. https://doi.org/10.1017/S0954394500001861.

Hickey, Raymond. (2002). Ebb and flow: A cautionary tale of language change. In Teresa Fanego, Belén Méndez-Naya & Elena Seoane (eds.), *Sounds, words, texts and change: Selected papers from 11 ICEHL, Santiago de Compostela, 7-11 September 2000* (volume 2), 105-128. John Benjamins Publishing Company.

Hodgson, Katherine. (2019). *Relative clauses in colloquial Armenian: Syntax and typology* [Unpublished doctoral dissertation]. INALCO. https://theses.hal.science/tel-02356401.

Hrach^{h}ya Acharyani anvan Lezvi institut [Language institute named after Hrach^{h}ya Acharyan]. (1977). *Hayereni barbaṛagitakan atlasi nyut^{h}eri havak^{h}man tsragir* [Program for the collection of materials of the dialectological atlas of Armenian]. Yerevan: Haykakan SSH GA Hratarakch^{h}ut^{h}yun [Publishing House of the Academy of Sciences of the Armenian SSR].

Irvine, Judith T. & Susan Gal. (2000). Language ideology and linguistic differentiation. In Paul V. Kroskrity (ed.), *Regimes of language: Ideologies, polities, and identities*, 35-84. Santa Fe: School of American Research Press.

Jahukyan, Gevorg Beglari. (1972). *Hay barbaṛagitut^{h}yan neratsut^{h}yun: Vichakagrakan barbaṛagitut^{h}yun* [Introduction to Armenian dialectology: Statistical dialectology]. Yerevan: Haykakan SSH GA Hratarakch^{h}ut^{h}yun [Publishing House of the Academy of Sciences of the Armenian SSR].

Janson, Tore. (1983). Sound change in perception and production. *Language* 59(1). 18-34. https://www.jstor.org/stable/414059.
Janson, Tore & Richard Schulman. (1983). Non-distinctive features and their use. *Journal of Linguistics* 19(2). 321-336. https://www.jstor.org/stable/4175696.
Johnson, Daniel Ezra. (2010). *Stability and change along a dialect boundary: The low vowels of southeastern New England* (Publication of the American Dialect Society 95). Durham, NC: Duke University Press.
Johnson, Keith, Edward Flemming, & Richard Wright. (1993). The Hyperspace Effect: Phonetic Targets Are Hyperarticulated. *Language* 69(3). 505-528. https://www.jstor.org/stable/416697.
Johnson, Keith, Elizabeth A. Strand, & Mariapaola D'Imperio. (1999). Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics* 27. 359-384. https://doi.org/10.1006/jpho.1999.0100.
Juskan, Marten Michael. (2016). *Production and perception of local variants in Liverpool English: Change, salience, exemplar priming* [Unpublished doctoral dissertation]. Freiburg University.
Kang, Shinae, Keith Johnson, Gregory Finley. (2016). Effects of native language on compensation for coarticulation. *Speech Communication* 77. 84-100. http://dx.doi.org/10.1016/j.specom.2015.12.005.
Karapetian, Shushan. (2014). *"How do I teach my kids my broken Armenian?": A study of Eastern Armenian heritage language speakers in Los Angeles.* Los Angeles: University of California, Los Angeles dissertation. https://escholarship.org/uc/item/7jq085nr.
Katvalyan, Viktor. (2009). *K^{h}äväṛa khänäk^{h}-ch^{h}änäk^{h} (Zvarchapatumner Bayazeti barbaṛov)* [Gavaṛ jokes (Anecdotes in the dialect of Bayazet)]. Yerevan: "Hayastan" hratarakch^{h}ut^{h}yun ["Armenia" Publishing House].
Katvalyan, Viktor. (2012). "A" dzaynavori drsevorumnerĕ Bayazeti barbaṛum ev "Achaṛyani orenk^{h}ĕ" [Manifestations of the vowel "a" in the Bayazet dialect and Achaṛyan's law]. *Banber Erevani hamalsarani. Banasirut^{h}yun* [Bulletin of Yerevan University: Philology] 138(2). 48-63.
Katvalyan, Viktor. (2016). *Bayazeti barbaṛĕ ev nra lezvakan aṛnch^{h}ut^{h}yunnerĕ shrjaka barbaṛneri het [The dialect of Bayazet and its linguistic connections with the surrounding dialects]*. Yerevan: "Asoghik" hratarakch^{h}ut^{h}yun ["Singer" Publishing House].
Katvalyan, Viktor. (2018). *Hayastani Hanrapetut^{h}yan barbaṛayin hamapatker: Gegharkh^{h}unik^{h}i marz* [Dialectal overview of the Republic of Armenia: Region of Gegharkh^{h}unik^{h}] (vol. 1). Yerevan: "Asoghik" hratarakch^{h}ut^{h}yun ["Singer" Publishing House].
Katvalyan, Viktor. (2021). *Bayazeti barbaṛi baṛaran* [Dictionary of the Bayazet dialect]. Yerevan: HH GAA LI hratarakch^{h}ut^{h}yun [Publishing House of the Institute of Language of the National Academy of Sciences of the Republic of Armenia].
Katvalyan, Viktor L. (2025). Dialectological atlas of the Armenian language: Part A. Gegharkunik region. Yerevan: Institute of Language of the NAS of the Republic of Armenia.
Katvalyan, V. L. & A.A. Abrahamyan (eds.). (2016). *Hayereni barbaṛagitakan atlas: Prak 4* [Dialectological atlas of Armenian: Part 4]. Yerevan: "Asoghik" hratarakch^{h}ut^{h}yun ["Singer" Publishing House].
Katvalyan, V., H. Khach^{h}atryan & G. Mkrtch^{h}yan (2021). *Hayastani Hanrapetut^{h}yan barbaṛayin hamapatker. Nyut^{h}er. Prak A. Gegharkh^{h}unik^{h}* [Dialectal overview of the Republic of

Armenia: Materials: Part A: Geghark^{h}unik^{h}]. Yerevan: GAA LI hratarakch^{h}ut^{h}yun [Publishing House of the Institute of Language of the National Academy of Sciences].
Kelley, Matthew C. & Benjamin V. Tucker. (2020). A comparison of four vowel overlap measures. *Journal of the Acoustical Society of America* 147. 137-145. https://doi.org/10.1121/10.0000494.
Kendall, Tyler & Erik R. Thomas. (2018). vowels: Vowel manipulation, normalization, and plotting. Version 1.2-2. https://CRAN.R-project.org/package=vowels.
Kendall, Tyler & Valerie Fridland. (2012). Variation in perception and production of mid front vowels in the U.S. Southern Vowel Shift. *Journal of Phonetics* 40. 289-306. https://doi.org/10.1016/j.wocn.2011.12.002.
Kendall, Tyler & Valerie Fridland. (2017). Regional relationships among the low vowels of U.S. English: Evidence from production and perception. *Language Variation and Change* 29. 245-271. https://doi.org/10.1017/S0954394517000084.
Kendall, Tyler & Valerie Fridland. (2021). *Sociophonetics*. Cambridge University Press.
Kendall, Tyler, Nicolai Pharao, Jane Stuart-Smith & Charlotte Vaughn. (2023). Advancements of phonetics in the 21st century: Theoretical issues in sociophonetics. *Journal of Phonetics* 98. https://doi.org/10.1016/j.wocn.2023.101226.
Kerswill, Paul. (2001). A dialect with 'great inner strength'? The perception of nativeness in the Bergen speech community. *Reading Working Papers in Linguistics* 5. 23-49.
Kettig, Thomas & Bodo Winter. (2017). Producing and perceiving the Canadian Vowel Shift: Evidence from a Montreal community. *Language Variation and Change* 29. 79-100. https://doi.org/10.1017/S0954394517000023.
Kewley-Port, Diane. (2001). Vowel formant discrimination II: Effects of stimulus uncertainty, consonantal context, and training. *Journal of the Acoustical Society of America* 110(4). 2141-2155. https://doi.org/10.1121/1.1400737.
Khach^{h}atryan, Amalya Harut^{h}yuni. (1988). *Zhamanakakits^{h} hayereni hnch^{h}uyt^{h}abanut^{h}yun [Phonology of modern Armenian]*. Haykakan SSH GA Hratarakch^{h}ut^{h}yun [Publishing House of the Academy of Sciences of the Armenian SSR].
Khachaturian, Amalia H. (1992). Voiced aspirated consonants in the Nor Bayazet dialect of Armenian. In John A.C. Greppin (ed.), *Proceedings of the fourth international conference on Armenian linguistics,* 115–128. Delmar, NY: Caravan Books.
Khachaturian, Amalia H. (1996). The Armenian sound system from typological aspect. In D. Sakayan (ed.), *Proceedings of the fifth international conference on Armenian linguistics*, 177-189. Caravan Books.
Kharlamov, Viktor. (2014). Incomplete neutralization of the voicing contrast in word-final obstruents in Russian: Phonological, lexical, and methodological influences. *Journal of Phonetics* 43. 47-56. https://doi.org/10.1016/j.wocn.2014.02.002.
Kirkham, Sam, Claire Nance, Bethany Littlewood, Kate Lightfoot, & Eva Groarke. (2019). Dialect variation in formant dynamics: The acoustics of lateral and vowel sequences in Manchester and Liverpool English. *Journal of the Acoustical Society of America* 145. 784-794. https://doi.org/10.1121/1.5089886.
Kleber, Felicitas, Jonathan Harrington, & Ulrich Reubold. (2012). The Relationship between the perception and production of coarticulation during a sound change in progress. *Language and Speech* 55(3). 383-405. https://doi.org/10.1177/0023830911422194.

Kleber, Felicitas, Tina John, & Jonathan Harrington. (2010). The implications for speech perception of incomplete neutralization of final devoicing in German. *Journal of Phonetics* 38. 185-196. https://doi.org/10.1016/j.wocn.2009.10.001.
Kolinsky, Régine, Ana Luiza Navas, Fraulein Vidigal de Paula, Nathalia Ribeiro de Brito, Larissa de Medeiros Botecchia, Sophie Bouton & Willy Serniclaes. (2021). The impact of alphabetic literacy on the perception of speech sounds. *Cognition* 213. 104687. https://doi.org/10.1016/j.cognition.2021.104687.
König, E. (1957). Pitch discrimination and age. *Acta Oto-Laryngologica* 48(5-6). 475-489. https://doi.org/10.3109/00016485709126909.
Kontra, Miklós. (1993). The messy phonology of Hungarians in South Bend: A contribution to the study of near-mergers. *Language Variation and Change* 5. 225-231. https://doi.org/10.1017/S0954394500001472.
Koops, Christian, Elizabeth Gentry, & Andrew Pantos. (2008). The effect of perceived speaker age on the perception of PIN and PEN vowels in Houston, Texas. *University of Pennsylvania Working Papers in Linguistics* 14(2). 93-101.
Korner-Nievergelt, Franzi, Tobias Roth, Stefanie von Felten, Jérôme Guélat, Bettina Almasi & Pius Korner-Nievergelt. (2015). *Bayesian data analysis in ecology using linear models with R, BUGS, and Stan*. Elsevier Science & Technology.
Kotsoni, Zoi. (2019). *A sociophonetic analysis of the production of mid-vowel contrasts in Catalan spoken in Barcelona* [Unpublished doctoral dissertation]. Universitat Autònoma de Barcelona.
Kraljic, Tanya, Susan E. Brennan, & Arthur G. Samuel. (2008). Accommodating variation: Dialects, idiolects, and speech processing. *Cognition* 107. 54-81. https://doi.org/10.1016/j.cognition.2007.07.013.
Kristiansen, Gitte. (2003). How to do things with allophones: Linguistic stereotypes as cognitive reference points in social cognition. In René Dirven, Roslyn Frank, Martin Pütz, René Dirven, & Martin Pütz (eds.), *Cognitive models in language and thought: Ideology, metaphors and meanings*, 69-120. Berlin: Mouton de Gruyter.
Kuang, Jianjing & Aletheia Cui. (2018). Relative cue weighting in production and perception of an ongoing sound change in Southern Yi. *Journal of Phonetics* 71. 194-214. https://doi.org/10.1016/j.wocn.2018.09.002.
Kuhl, Patricia K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics* 50(2). 93-107. https://doi.org/10.3758/bf03212211.
Kutlu, Ethan, Keith Baxelbaum, Eldon Sorensen, Jacob Oleson & Bob McMurray. (2024). Linguistic diversity shapes flexible speech perception in school age children. *Scientific Reports* 14. 28825. https://doi.org/10.1038/s41598-024-80430-1.
Kuznetsova Alexandra, Per B. Brockhoff & Rune H.B. Christensen. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* 82(13). 1-26. https://doi.org/10.18637/jss.v082.i13.
Labov, William. (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.
Labov, William. (1994). *Principles of linguistic change: Internal factors*. Oxford: Blackwell.
Labov, William. (2001). *Principles of linguistic change: Social factors.* Oxford: Blackwell.
Labov, William. (2010). *Principles of linguistic change: Cognitive and cultural factors*. Chichester, UK: Wiley-Blackwell.

Labov, William, Sharon Ash, & Charles Boberg. (2006). *The atlas of North American English: Phonetics, phonology, and sound change*. De Gruyter Mouton.
Labov, William, Sabriya Fisher, Duna Gylfadottír, Anita Henderson & Betsy Sneller. (2016). Competing systems in Philadelphia phonology. *Language Variation and Change* 28. 273-305. https://doi.org/10.1017/S0954394516000132.
Labov, William, Mark Karen, & Corey Miller. (1991). Near-mergers and the suspension of phonemic contrast. *Language Variation and Change* 3. 33-74. https://doi.org/10.1017/S0954394500000442.
Labov, William, Malcah Yaeger, & Richard Steiner. (1972). *A quantitative study of sound change in progress.* The U.S. Regional Survey.
Lee, Jinsok. (2016). The Participation of a northern New Jersey Korean American Community in local and national language variation. *American Speech* 91(3). 327-360. https://doi.org/10.1215/00031283-3701037.
Lenth, Russell V. & Julia Piaskowski. (2026). *emmeans: Estimated marginal means, aka least-squares means*. R package version 2.0.3. https://github.com/rvlenth/emmeans.
Liberman, Alvin M., Katherine Safford Harris, Howard S. Hoffman, and Belver C. Griffith. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology* 54(5). 358-368. https://doi.org/10.1037/h0044417.
Loakes, Deborah, Joshua Clothier, John Hajek, & Janet Fletcher. (2014). An investigation of the /el/–/æl/ merger in Australian English: A pilot study on production and perception in South-West Victoria. *Australian Journal of Linguistics* 34(4). 436-452. https://doi.org/10.1080/07268602.2014.929078.
Mack, Sara & Benjamin Munson. (2012). The influence of /s/ quality on ratings of men's sexual orientation: Explicit and implicit measures of the 'gay lisp' stereotype. *Journal of Phonetics* 40. 198-212. https://doi.org/10.1016/j.wocn.2011.10.002.
Macmillan, Neil A. & C. Douglas Creelman (2005). *Detection theory: A user's guide* (2nd edn.). New York: Psychology Press.
Maguire, Warren Noel. (2007). What is a merger, and can it be reversed? The origin, status and reversal of the 'NURSE-NORTH merger' in Tyneside English [Unpublished doctoral dissertation]. University of Newcastle Upon Tyne.
Maguire, Warren, Lynn Clark, & Kevin Watson. (2013). Introduction: What are mergers and can they be reversed? *English Language and Linguistics* 17(2). 229-239. https://doi.org/10.1017/S1360674313000014.
Mann, Virginia. (1980). Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics* 28(5). 407-412. https://doi.org/10.3758/BF03204884.
Mann, Virginia A. & Bruno H. Repp. (1980). Influence of vocalic context on perception of the [ʃ]-[s] distinction. *Perception & Psychophysics* 28(3). 213-228. https://doi.org/10.3758/BF03204377.
Mann, Virginia A. & Bruno H. Repp. (1981). Influence of preceding fricative on stop consonant perception. *Journal of the Acoustical Society of America* 69. 548-558. https://doi.org/10.1121/1.385483.
Manoukian, Jennifer. (2023). *In search of purity: Language, ideology and global intellectual movements in Ottoman Armenian history, 1750-1915* [Unpublished doctoral dissertation]. University of California, Los Angeles. https://escholarship.org/uc/item/37z8s351.
Markosyan, Ṛazmik. (1989). *Araratyan barbaṛ* [Ararat dialect]. Yerevan: Luys [Light].

Masapollo, Matthew, Linda Polka, Monika Molnar, & Lucie Ménard. (2017). Directional asymmetries reveal a universal bias in adult vowel perception. *Journal of the Acoustical Society of America* 141. 2857-2869. https://doi.org/10.1121/1.4981006.
Mazerolle, MJ. (2023). *AICcmodavg: Model selection and multimodel inference based on (Q)AIC(c)*. R package version 2.3.3. https://cran.r-project.org/package=AICcmodavg.
McGowan, Kevin B. (2015). Social Expectation Improves Speech Perception in Noise. *Language and Speech* 58(4). 502-521. https://doi.org/10.1177/0023830914565191.
McGowan, Kevin B. & Anna M. Babel. (2019). Perceiving isn't believing: Divergence in levels of sociolinguistic awareness. *Language in Society* 49. 231-256.
McGuire, Grant. (2010). A brief primer on experimental designs for speech perception research [Unpublished manuscript]. UC Santa Cruz. https://scispace.com/pdf/el-diptongo-hiato-como-rasgo-contrastivo-un-estudio-kwia99hlgt.pdf.
McMurray, Bob. (2022). The myth of categorical perception. *Journal of the Acoustical Society of America* 152(6). https://doi.org/10.1121/10.0016614.
MediaHub. (2024). [Photograph of road sign in Gavaṛ]. https://mediahub.am/post/ee35108e097ed119.
Mielke, Jeff, Christopher Carignan, & Erik R. Thomas. (2017). The articulatory dynamics of pre-velar and pre-nasal /æ/-raising in English: An ultrasound study. *Journal of the Acoustical Society of America* 142. 332-349. https://doi.org/10.1121/1.4991348.
Milroy, James & Lesley Milroy (1978). Belfast: change and variation in an urban vernacular. In Peter Trudgill (ed.), *Sociolinguistic patterns in British English*, 19-36. London: Edward Arnold.
Milroy, James & Lesley Milroy. (2012). *Authority in Language: Investigating Standard English* (4th ed.). Routledge.
Mitterer, Holger. (2006). On the causes of compensation for coarticulation: Evidence for phonological mediation. *Perception & Psychophysics* 68(7). 1227-1240. https://doi.org/10.3758/BF03193723.
Mitterer, Holger & Mirjam Ernestus. (2008). The link between speech perception and production is phonological and abstract: Evidence from the shadowing task. *Cognition* 109. 168-173. http://doi.org/10.1016/j.cognition.2008.08.002.
Miyawaki, Kuniko, Winifred Strange, Robert Verbrugge, Alvin M. Liberman, James J. Jenkins & Osamu Fujimura. (1975). An effect of linguistic experience: The discrimination of [r] and [1] by native speakers of Japanese and English. *Perception & Psychophysics* 18(5). 331-340.
Mkrtch^{h}yan, Garik. G. (2015). Gegharkʰunikʰi marzi khosvatskʰneri baghadzaynneri taratskʰayin bnutʰagirĕ [Territorial characterization of the consonants of dialects of Gegharkʰunikʰ region]. *Lezu ev Lezvabanutʰyun [Language and Linguistics]* 1-2. 53-63. https://arar.sci.am/dlibra/publication/216328/edition/197068/content.
Mkrtch^{h}yan, Garik. G. (2016). Gegharkʰunikʰi marzi khosvatskʰneri dzaynavorneri ev erkbarbaṛneri taratskʰayin bnutʰagirĕ [Territorial characterization of the vowels and diphthongs of dialects of Gegharkʰunikʰ region]. *Lraber hasarakakan gitutʰyunneri* [Herald of the Social Sciences] 3(3). 199-207. https://arar.sci.am/dlibra/publication/44464/edition/39901/content.
Morgan-Wall, Tyler. (2020). Rayshader: Create Maps and Visualize Data in 2D and 3D. R package version 0.19.1. http://www.rayshader.com.

Morris, Midori Yonezawa. (2010). Regional Stereotypes and the Perception of Japanese Vowel Devoicing. In Dennis Preston & Nancy Niedzielski (eds.), *A reader in sociophonetics*, 191-202. De Gruyter Mouton.

Munson, Benjamin. (2011). The influence of actual and imputed talker gender on fricative perception, revisited (L). *Journal of the Acoustical Society of America* 130(5). 2631-2634. https://doi.org/10.1121/1.3641410.

Munson, Benjamin, Kayleigh Ryherd, & Sara Kemper. (2017). Implicit and explicit gender priming in English lingual sibilant fricative perception. *Linguistics* 55(5). 1073-1107. https://doi.org/10.1515/ling-2017-0021.

Nahhas, Ramzi W. (2025). *Introduction to Regression Methods for Public Health Using R*. Chapman & Hall/CRC Press. https://019b2da8-edfb-a262-61be-7973c056d9ae.share.connect.posit.cloud/mlr-collinearity.html.

Newman, Rochelle S. (2003). Using links between speech perception and speech production to evaluate different acoustic metrics: A preliminary report. *Journal of the Acoustical Society of America* 113. 2850–2860. https://doi.org/10.1121/1.1567280.

Nicenboim, Bruno, Timo B. Roettger, & Shravan Vasishth. (2018). Using meta-analysis for evidence synthesis: The case of incomplete neutralization in German. *Journal of Phonetics* 70. 39-55. https://doi.org/10.1016/j.wocn.2018.06.001.

Niedzielski, Nancy. (1999). The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology* 18(1). 62-85. https://doi.org/10.1177/0261927X99018001005.

Nunberg, Geoffrey. (1980). A falsely reported merger in eighteenth-century English: A study in diachronic variation. In William Labov (ed.), *Locating language in time and space*, 221-250. New York: Academic Press.

Nycz, Jennifer R. (2011). *Second dialect acquisition: Implications for theories of phonological representation* [Unpublished doctoral dissertation]. New York University.

Nycz, Jennifer & Lauren Hall-Lew. (2014). Best practices in measuring vowel merger. *Proceedings of Meetings on Acoustics: The Acoustical Society of America through the American Institute of Physics* 20(1). 1-19. https://doi.org/10.1121/1.4894063.

O'Callaghan, Alan, Max Kuhn, Steve Weston, Jed Wing, James Forester & Thorn Thaler. (2022). *A collection of contrast methods.* https://cran.r-project.org/web/packages/contrast/contrast.pdf.

Oh, Sejin. (2022). *Incomplete neutralization in articulatory phonology* [Unpublished doctoral dissertation]. The City University of New York.

Ohala, John J. (1981). The Listener as a Source of Sound Change. In Carrie S. Masek, Roberta A. Hendrick & Mary F. Miller (eds.), *Papers from the parasession on language behavior*, 178-203. Chicago Linguistic Society.

Oshagan, Vahé. (1997). Modern Armenian literature and intellectual history from 1700 to 1915. In Richard G. Hovannisian (ed.), *Foreign dominion to statehood: The fifteenth century to the twentieth century* (Vol. 2 in Richard G. Hovannisian (ed.), *The Armenian people from ancient to modern times*), 135-174. Houndmills: Macmillan.

Paliwal, K.K., D. Lindsay, & W.A. Ainsworth. (1983). Correlation between production and perception of English vowels. *Journal of Phonetics* 11. 77-83. https://doi.org/10.1016/S0095-4470(19)30778-8.

Pardoe, Iain. (2020). *Applied regression modeling* (3rd edn.). Hoboken, NJ: John Wiley & Sons.

Peirce, Jonathan, Jeremy R. Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyuki Sogo, Erik Kastman, & Jonas Kristoffer Lindeløv. (2019). PsychoPy2: experiments in behavior made easy. *Behavior Research Methods* 51. 195-203. https://doi.org/10.3758/s13428-018-01193-y.

Peperkamp, Sharon, Michèle Pettinato, & Emmanuel Dupoux. (2003). Allophonic variation and the acquisition of phoneme categories. In Barbara Beachley, Amanda Brown & Frances Conlin (eds.), *Proceedings of the 27th annual Boston University conference on language development, volume 2*, 650–661. Sommerville, MA: Cascadilla Press.

Perkell, Joseph S., Frank H.Guenther, Harlan Lane, Melanie L. Matthies, & Ellen Stockmann. (2004). The distinctness of speakers' productions of vowel contrasts is related to their discrimination of the contrasts. *Journal of the Acoustical Society of America* 116(4). 2338-2344. https://doi.org/10.1121/1.1787524.

Peterson, Gordon E. & Harold L. Barney. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America, 24*(2), 175-184.

Petersen, Jan Heegård. (2024). Enhanced coarticulatory labialization of /tˢ/ in Argentine Danish. *Nordic Journal of Linguistics*. 1-27. https://doi.org/10.1017/S0332586524000040.

Pinget, Anne-France., René Kager, & Hans Van de Velde. (2020). Linking variation in perception and production in sound change: Evidence from Dutch obstruent devoicing. *Language and Speech* 63(3). 660-685. https://doi.org/10.1177/0023830919880206.

Port, Robert & Penny Crawford. (1989). Incomplete neutralization and pragmatics German. *Journal of Phonetics* 17. 257-282. https://doi.org/10.1016/S0095-4470(19)30444-9.

Portugal, Emma. (2025). Dialect vowel maintenance in a diglossic context: a study of linguistic heterogeneity in Gavar, Armenia. *Folia Linguistica* 60(1). 391-432. https://doi.org/10.1515/flin-2025-2019.

Portugal, Emma & Sean Nonnenmacher. (2024). "Every word is a world": Loanword ideologies and linguistic purism in post-Soviet Armenia. *Multilingua* 43(3). 331-364. https://doi.org/10.1515/multi-2022-0152.

Preston, Dennis R. (2016). Whaddayaknow now? In Anna M. Babel (ed.), *Awareness and control in sociolinguistic research*, 177-199. Cambridge University Press.

R Core Team. (2021). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/.

Rathcke, Tamara, Jane Stuart-Smith, Bernard Torsney & Jonathan Harrington. (2017). The beauty in a beast: Minimising the effects of diverse recording quality on vowel formant measurements in sociophonetic real-time studies. *Speech Communication* 86. 24-41. https://dx.doi.org/10.1016/j.specom.2016.11.001.

Recasens, Daniel & Aina Espinosa. (2009). Dispersion and variability in Catalan five and six peripheral vowel systems. *Speech Communication* 51. 240-258. https://doi.org/10.1016/j.specom.2008.09.002.

Renwick, Margaret E. (2017). Dynamic trajectories of tense vs. lax vowels in the American South. *Journal of the Acoustical Society of America* 142. 2680. https://doi.org/10.1121/1.5014773.

Renwick, Margaret E.L. & Joseph A. Stanley. (2020). Modeling dynamic trajectories of front vowels in the American South. *Journal of the Acoustical Society of America* 147. 579-595. https://doi.org/10.1121/10.0000549.

Repp, Bruno H. & Virginia A. Mann. (1981). Perceptual assessment of fricative–stop coarticulation. *Journal of the Acoustical Society of America* 69. 1154-1163. https://doi.org/10.1121/1.385695.
Roettger, T.B., B. Winter, S. Grawunder, J. Kirby, & M. Grice. (2014). Assessing incomplete neutralization of final devoicing in German. *Journal of Phonetics* 43. 11-25. https://doi.org/10.1016/j.wocn.2014.01.002.
Rose, Sharon, Michael Obiri-Yeboah, & Sarah C. Creel. (2023). Perception of ATR contrasts by Akan speakers: a case of perceptual near-merger. *Laboratory Phonology* 14(1). 1-39. https://doi.org/10.16995/labphon.8948.
Schertz, Jessamyn, Yoonjung Kang, & Sungwoo Han. (2019). Sources of variability in phonetic perception: The joint influence of listener and talker characteristics on perception of the Korean stop contrast. *Laboratory Phonology* 10(1). 1-32. https://doi.org/10.5334/labphon.67.
Seyfarth, Scott, Hossep Dolatian, Peter Guekguezian, Niamh Kelly & Tabita Toparlak. (2023). Armenian (Yerevan Eastern Armenian and Beirut Western Armenian). *Journal of the International Phonetic Association*. 1-34. https://doi.org/10.1017/S0025100323000130.
Seyfarth, Scott & Marc Garellek. (2018). Plosive voicing acoustics and voice quality in Yerevan Armenian. *Journal of Phonetics* 71. 425-450. https://doi.org/10.1016/j.wocn.2018.09.001.
Sloos, Marjoleine. (2013). The reversal of the Bären–Beeren merger in Austrian Standard German. *The Mental Lexicon* 8(3). 353-371. https://doi.org/10.1075/ml.8.3.05slo.
Sóskuthy, Márton. (2017). Generalised additive mixed models for dynamic analysis in linguistics: A practical introduction. https://eprints.whiterose.ac.uk/113858/2/1703_05339v1.pdf.
Squires, Lauren. (2013). It don't go both ways: Limited bidirectionality in sociolinguistic perception. *Journal of Sociolinguistics* 17(2). 200-237. https://doi.org/10.1111/josl.12025.
Squires, Lauren. (2016). Processing Grammatical Differences: Perceiving versus Noticing. In Anna M. Babel (ed.), *Awareness and control in sociolinguistic research*, 80-103. Cambridge University Press.
Stanley, Joey. (2020). Barktools. https://joeystanley.github.io/barktools/.
Stanley, Joey. (2021, February 6). *Pillai scores don't change after normalization.* Joey Stanley, assistant professor, linguistics. https://joeystanley.com/blog/pillai-scores-dont-change-after-normalization/.
Stanley, Joseph A. & Betsy Sneller. (2023). Sample size matters in calculating Pillai scores. *Journal of the Acoustical Society of America* 153(1). 54-67. https://doi.org/10.1121/10.0016757.
Statistical committee of the Republic of Armenia. (2022). The Main Results of RA Census 2022. Armstat. https://www.armstat.am/en/?nid=82&id=2623.
Steinberg, Johanna, Hubert Truckenbrodt, & Thomas Jacobsen. (2012). The role of stimulus cross-splicing in an event-related potentials study. Misleading formant transitions hinder automatic phonological processing. *Journal of the Acoustical Society of America* 131. 3120-3140. https://doi.org/10.1121/1.3688515.
Strand, Elizabeth. (1999). Uncovering the role of gender stereotypes in speech perception. *Journal of Language and Social Psychology* 18(1). 86-100. https://doi.org/10.1177/0261927X99018001006.

Strandberg, Janine A.E., Charlotte Gooskens & Anja Schüppert. (2022). Re-examining the /eː-ɛː/ merger in Finland-Swedish: Regional and stylistic variation. *Language Variation and Change* 34. 183-214. https://doi.org/10.1017/S0954394522000072.

Strelluf, Christopher. (2014). *"We have such a normal, non-accented voice": A sociophonetic study of English in Kansas City* [Unpublished doctoral dissertation]. University of Missouri-Columbia.

Suhas, Satish, Narayana Manjunatha, Channaveerachari Naveen Kumar, Vivek Benegal, Girish N. Rao, Mathew Varghese & Gopalkrishna Gururaj. (2023). Firth's penalized logistic regression: A superior approach for analysis of data from India's National Mental Health Survey, 2016. *Indian Journal of Psychiatry* 65(12). 1208-1213. https://doi.org/10.4103/indianjpsychiatry.indianjpsychiatry_827_23.

Sumner, Meghan & Arthur G. Samuel. (2009). The effect of experience on the perception and representation of dialect variants. *Journal of Memory and Language* 60. 487-501. https://doi.org/10.1016/j.jml.2009.01.001.

TheGlobalEconomy.com. (2023). *Armenia: Female to male ratio, students at tertiary level education.* https://www.theglobaleconomy.com/Armenia/Female_to_male_ratio_students_tertiary_level_educa/.

Thomas, Brynmor & Jennifer Hay. (2005). A pleasant malady: The Ellen/Allan merger in New Zealand English. *Te Reo* 48. 69-93.

Thomas, Erik R. (2000). Spectral differences in /ai/ offsets conditioned by voicing of the following consonant. *Journal of Phonetics* 28. 1-25. https://doi.org/10.1006/jpho.2000.0103.

Thomas, Erik R. (2002). Sociophonetic applications of speech perception experiments. *American Speech* 77(2). 115-147. https://muse.jhu.edu/article/2853.

Thomas, Erik R. (2011). *Sociophonetics: An introduction*. Basingstoke, UK: Palgrave Macmillan.

Trudgill, Peter. (1986). *Dialects in contact*. Oxford: Basil Blackwell.

Vaux, Bert. (1998). *The phonology of Armenian*. Oxford: Oxford University Press.

Viswanathan, Navin, James S. Magnuson, & Carol A. Fowler. (2010). Compensation for coarticulation: Disentangling auditory and gestural theories of perception of coarticulatory effects in speech. *Journal of Experimental Psychology: Human Perception and Performance* 36(4). 1005-1015. https://doi.org/10.1037/a0018391.

Wagner, Anita, Mirjam Ernestus, & Anne Cutler. (2006). Formant transitions in fricative identification: The role of native fricative inventory. *Journal of the Acoustical Society of America* 120. 2267-2277. https://doi.org/10.1121/1.2335422.

Warburton, Jasmine. (2020). *The merging of the goat and thought vowels in Tyneside English: Evidence from production and perception* [Unpublished doctoral dissertation]. Newcastle University.

Warner, Natasha, Allard Jongman, Joan Sereno, & Rachel Kemps. (2004). Incomplete neutralization and other sub-phonemic durational differences in production and perception: evidence from Dutch. *Journal of Phonetics* 32. 251-276. https://doi.org/10.1016/S0095-4470(03)00032-9.

Weitenberg, J.J.S. (2002). Aspects of Armenian dialectology. In Jan Berns & Jaap van Marle (eds.), *Present-day dialectology: Problems and findings*, 141-157. Berlin: Mouton de Gruyter.

Wickham, Hadley. (2016). ggplot2: Elegant graphics for data analysis. New York: Springer-Verlag. https://ggplot2.tidyverse.org.
Winn, Matthew. (2023). Praat Scripts. http://www.mattwinn.com/praat.html#formantContinuum.
Wong, Amy Wing-mei & Lauren Hall-Lew. (2014). Regional variability and ethnic identity: Chinese Americans in New York City and San Francisco. *Language & Communication* 35. 27-42. http://dx.doi.org/10.1016/j.langcom.2013.11.003.
Xu, Robert. (2019). Ongoing change of PIN/PEN vowels in California's Central Valley. *International Congress of Phonetic Sciences (ICPhS 2019).* https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2019/papers/ICPhS_511.pdf.
Zakarian, Hovhannes L. (1996). The language law of the Republic of Armenia and problems of all-Armenian language policy. In Dora Sakayan (ed.), *Proceedings of the fifth international conference on Armenian linguistics*, 177-189. Delmar, NY: Caravan Books.
Zellou, Georgia. (2017). Individual differences in the production of nasal coarticulation and perceptual compensation. *Journal of Phonetics* 61. 13-29. http://dx.doi.org/10.1016/j.wocn.2016.12.002.
Zhang, Jennifer, Lindsey Graham, Marissa Barlaz, & José Ignacio Hualde. (2022). Within-speaker perception and production of two marginal contrasts in Illinois English. *Frontiers in Communication* 7. https://doi.org/10.3389/fcomm.2022.844862.