

INFORMATION TO USERS

This material was produced from a microfilm copy of the original document. While the most advanced technological means to photograph and reproduce this document have been used, the quality is heavily dependent upon the quality of the original submitted.

The following explanation of techniques is provided to help you understand markings or patterns which may appear on this reproduction.

- 1. The sign or "target" for pages apparently lacking from the document photographed is "Missing Page(s)". If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting thru an image and duplicating adjacent pages to insure you complete continuity.**
- 2. When an image on the film is obliterated with a large round black mark, it is an indication that the photographer suspected that the copy may have moved during exposure and thus cause a blurred image. You will find a good image of the page in the adjacent frame.**
- 3. When a map, drawing or chart, etc., was part of the material being photographed the photographer followed a definite method in "sectioning" the material. It is customary to begin photoing at the upper left hand corner of a large sheet and to continue photoing from left to right in equal sections with a small overlap. If necessary, sectioning is continued again — beginning below the first row and continuing on until complete.**
- 4. The majority of users indicate that the textual content is of greatest value, however, a somewhat higher quality reproduction could be made from "photographs" if essential to the understanding of the dissertation. Silver prints of "photographs" may be ordered at additional charge by writing the Order Department, giving the catalog number, title, author and specific pages you wish reproduced.**
- 5. PLEASE NOTE: Some pages may have indistinct print. Filmed as received.**

Xerox University Microfilms

300 North Zeeb Road
Ann Arbor, Michigan 48106

76-27,449

BERNSTEIN, Jared Christopher, 1948-
VOCOID PSYCHOACOUSTICS, ARTICULATION
AND VOWEL PHONOLOGY.

The University of Michigan, Ph.D., 1976
Psychology, experimental

Xerox University Microfilms, Ann Arbor, Michigan 48106

VOCOID PSYCHOACOUSTICS, ARTICULATION
AND VOWEL PHONOLOGY

by
Jared Christopher Bernstein

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Psycholinguistics)
in The University of Michigan
1976

Doctoral Committee:

Professor John C. Catford, Co-Chairman
Professor David H. Krantz, Co-Chairman
Professor Kenneth L. Pike
Professor Irwin Pollack
Dennis H. Klatt, Research Associate, M.I.T.

به مارگرت قربانت

ACKNOWLEDGEMENTS

First I would like to thank Kenneth Pike and David Krantz. The standards they set in their own work and their advice and encouragement was a considerable help throughout my graduate career. They both show unusually warm concern for students as people.

I am also grateful to Irv Pollack for his fearless editing and to Dennis Klatt for his help at MIT and his comments on the completed manuscript. Kenneth Stevens' invitation to spend a year at MIT made this research possible. I appreciate the time he spent with me there as well as the patient assistance of Keith North, Bill Huggins, Shinji Maeda, and Bill Henke. I also had spiritual assistance from Steve Herman and Mike O'Malley.

Thanks to J. C. Catford's effective teaching, I had a fascinating and infectious introduction to his subject, phonetics.

Laura Smedley, who can type faster than I can read, is to be affectionately commended for her endurance in the face of unreason.

My final and special thanks are for Marie Bernstein, whose continuous moral and financial support long pre-dated my academic ambitions.

At last I'd like to apologize to Winky and Bongo and anyone else who is not ready for this tome. I mistook their confidence for unconcern.

TABLE OF CONTENTS

DECIDATION	ii
ACKNOWLEDGEMENTS	iii
LIST OF TABLES	v
CHAPTER ONE: INTRODUCTION	1
Projects	
Phonetics and Vocoids	
Phonetic Influence on Vowel Systems	
Project III Explained	
CHAPTER TWO: METHODOLOGICAL REVIEW	20
Language Independence in Perception	
Auditory Quality of Vocoid Transfer Functions	
Measurement Theory and Project I	
Measuring and Modeling the Vocal Tract	
CHAPTER THREE: AUDITORY TESTS	48
Acoustic Range of Possible Vocoids	
Theory	
General Experimental Method	
Experiment 1	
Experiment 2	
Experiment 3	
Experiment 4	
Results Summary	
CHAPTER FOUR: ARTICULATION AND DISTANCE	114
Digital Vocal Tract Model	
Oral Proprioception	
Conclusion	
CHAPTER FIVE: APPLICATIONS	129
Implementing Maxi	
Implementing QNS	
CHAPTER SIX: CAVEATS AND CONCLUSIONS	135
Caveats	
Conclusions	
APPENDIX: FIGURES	150
REFERENCES	188

LIST OF TABLES

3-1: Stimuli of Experiment 1	64
3-2: Frequency Intervals in Experiment 1	65
3-3a,b: Ordered Frequency Intervals	66
3-4: Results from Experiment 1	67
3-5: Orders of Intervals in Experiment 1	69
3-6: Results from Experiment 2a	75
3-7a,b: Definitions and Results of Experiment 2a	77
3-8: Distance Quadruples in Experiment 2a	78
3-9: Stimuli in Second Part of Experiment 2b	84
3-10: Data from Experiment 2b	85
3-11: Stimuli of Experiment 3	92
3-12: Data from Experiment 3	93
3-13: Stimuli for Experiment 4	99
3-14: Results from Experiment 4	100
3-15: F1 and F2 Intervals Ordered	101

CHAPTER ONE

INTRODUCTION

Scholars often explain phonological processes and structures with reference to general principles of phonetics or psychology. In this manner, phonological structures reflected in sound change, environmental conditioning, phonotactics, or in inventories of contrasts or segments are explained in extra-linguistic terms such as phonetic possibility, ease, similarity, or symmetry. The present work describes the construction of a quantitative model of the phonetic medium for the transmission of a restricted set of sounds and reports the results of applying this model to predict segment inventories.

This first chapter introduces an approach to predicting vowel inventories (section 1.1) and supplies background material on vocoids and their acoustic properties (1.2). The relative importance of different stages in the phonetic channel is discussed in section 1.3, and two hypotheses relating 'ease' to vowel segment inventories are explained in section 1.4.

1.1 Projects.

The object of the research program reported here was to complete three related projects. The general goal was to construct a model of the phonetic channel through which vocoids are transmitted and to relate this model to vowel phonology. Figure 1.1-1 schematizes the phonetic channel as relevant to transmitting static vocoids. The figure represents a compound function by which scaled articulatory vectors, $\vec{P}$, are mapped into perceptual vectors, $\vec{\Psi}$, through two intermediate stages, $\vec{A}$ and $\vec{F}$. That is, the phonetic channel, as shown, takes an articulatory position $\vec{A} = (a_1, a_2, \dots, a_n)$ and translates it into an F-pattern, $\vec{F} = (f_1, f_2, f_3, f_4)$, as $G(\vec{A}) = \vec{F}$, and the F-pattern, $\vec{F}$, is translated into a perceptual vector, $\vec{\Psi} = (\psi_1, \dots, \psi_m)$, as $H(\vec{F}) = \vec{\Psi}$. Articulatory positions, $\vec{A}$, are shown in a one-to-one relation with their perceived counterparts for the speaker, $\vec{P}$, as $F(\vec{P}) = \vec{A}$ and $F^{-1}(\vec{A}) = \vec{P}$. The phonetic channel, then, is a function $H(G(F(\vec{P}))) = \vec{\Psi}$.

Project I experimentally identified a set of auditory dimensions for vocoids $\psi_1, \dots, \psi_m$ and determined function H , their psychophysical relation to the formant frequencies. Also a distance function, $d_\psi(v_i, v_j)$, on pairs of vocoids in the auditory domain was developed positing a combination rule for the auditory dimensions and scaling them to a common unit. Chapter 3 and most of Chapter 2 concern Project I.

Project II was the construction of functions G and F and a proprioceptive distance $d_p(v_i, v_j)$. Function G was realized as a model of the vocal tract mapping vocoids, $\bar{A}$, in articulatory space into points, $\bar{F}$, in formant space. Function F was to scale the articulatory positions, $\bar{A}$, to a speaker's proprioceptive sense, $\bar{P}$, of those positions on the basis of a small experiment that collected articulatory matches to points on psychological continua. Because the results of the matching experiment did not reveal any important proprioceptive distortions, articulatory distance for the speaker was ultimately implemented as a quasiphysical distance $d_a(v_i, v_j)$ on the parameters of the vocal tract model. Project II is described in Chapter 4.

Project III was the application of the quantitative phonetic relations $H(G(\bar{A}))$ and $d_a(v_i, v_j)$ and $d_\psi(v_i, v_j)$ in working through two meta-linguistic principles which hypothesize well defined operations within a phonetic system for predicting properties of phonological systems on the basis of least effort for the speaker and listener. These two operational definitions of 'ease' were applied to produce optimal vowel quality systems within the phonetic channel as developed in Projects I and II. Project III is explained in section 1.4, worked through in Chapter 5, and evaluated in Chapter 6.

1.2 Phonetics and Vocoids

This section will provide some background material for the general reader. Section 1.2.1 reviews the distinction between phonetics and phonology, and defines "vocoid" as distinguished from "vowel." Section 1.2.2 reviews the acoustic theory of ideal vocoids and identifies some pertinent analytical relations among acoustic properties of vocoids.

1.2.1 Vocoids vs Vowels. A phonetic examination of speech sounds relates primarily to the sounds' physical, physiological, and auditory properties. Phonological study determines which phonetic properties of the sounds in a language are relevant and which are redundant; phonology groups sounds by functional criteria within and across the sound systems of particular languages.

Phonetic description and research (versus e.g., medical research) reflects an interest in speech as a carrier of linguistic information. Regardless of this, many phonetic aspects of speech are language-independent and can be regarded as properties of a biological mechanism immersed in air, so that some phonetic facts and relations can be understood in purely physical terms, some in solely biological terms, and some in psychophysical terms.

We distinguish the terms "vowel" and "vocoid." When discussing a sound from a phonetic perspective, "vocoid" will be used to identify a sound with certain physical characteristics. A vocoid is produced with no constriction in the vocal tract sufficiently small to cause a pressure build-up along the path of air movement. Vocoids have no frication noise during normal voiced air flow. The term "vowel" will be reserved for identifiable classes of sounds (usually vocoids) that typically function as syllabic nuclei in relation to the sound system of a language. Vocoids need not be vowels; note h, y, and w in English. Vowels need not be vocoids; a language can use [s] or trilled [r] as a syllabic nucleus. English uses [n] and [l] as syllabic nuclei.

Vocoids are defined by Pike (1943, p. 140-43) as the "central resonant orals." Vocoid articulations are central in that they do not have a complete obstruction in the center of the vocal tract as [l] has; vocoids are resonant in that the path of air escape is nowhere sufficiently constricted to cause frication; and vocoids are oral in that air is allowed to escape from the mouth. 'Vocoid' is synonymous with "non-consonantal" in the phonetic framework of Chomsky and Halle (1968, p. 302). Non-consonantal segments are produced without "a radical obstruction in the midsagittal region of

the vocal tract." Both of the terms, vocoid and non-consonantal, include articulatory positions that allow air to escape from the nose, however we will use the term 'vocoid' to mean "non-nasal vocoid" unless otherwise specified.

1.2.2 Vocoid Acoustics. A speech wave resulting from voiced vocoid articulation has many acoustic properties that might be audible. However, a number of these acoustic properties are analytically interdependent, although this inter-dependence does not rule out their potential auditory independence. As an analogy, suppose that the perception of rectangular shapes were at issue (Krantz and Tversky, 1975), and it was suggested that height, width, area, and height/width were all salient qualities of rectangular shape, although any two properties (e.g., height and area) determine all four. Similarly, in the literature on vocoid perception, authors have suggested many different attributes of vocoids to be important determinants of the resulting auditory impression, although all of these properties are derivable from a smaller subset of them within an acoustical theory of vocoid production (see Stevens and House, 1961).

The speech wave of a voiced vocoid can be viewed as the response of a vocal tract as it filters a voice source, thus allowing the specification of vocoid production in two parts, source properties and filter properties, which correspond to phonation and articulation as traditionally distinguished (see Catford, 1968). Usually, terms like "vowel quality" or "vowel color" are identified with the acoustic characteristics of a speech wave that are imposed by the supralaryngeal articulatory filter, in opposition to "voice quality" or "phonation quality" which refers to characteristics that depend on the source volume velocity waveform that enters the vocal tract at the glottis. Although there tend to be some correlations among acoustic characteristics of the source (like amplitude, fundamental frequency, and spectral shape), the source characteristics (associated with laryngeal activity) are largely independent of the filter characteristics (associated with articulatory activity). The filter

characteristics of any given vocal tract configuration are specified by means of the transfer function of that configuration. The transfer function is the ratio of the output spectrum to the input spectrum (i.e., the ratio of the Fourier transforms of the output and input waveforms). This transfer function determines the amplitude attenuation and phase shift for each sinusoidal frequency in the input. The transfer function for any vocal tract configuration can be described as the contributions of an infinite number of resonances and anti-resonances, each of which is characterized by its frequency, $f = j\omega/2\pi$, and a damping constant, $b = \sigma/\pi$. Since there are no anti-resonances in a vocoid transfer function, the filtering effects of a vocoid vocal tract on the amplitude and phase of an incoming signal from the larynx are completely determined at every frequency by the complex positions of its characteristic resonances. These resonances, or normal modes of vibration, are called "formants" and all the formant frequencies characteristic of a vocal tract at a given time are called the "F-pattern." The F-pattern is the pivotal representation of vocoids that connects Projects I and II.

Keeping constant the spectral properties of a source of excitation to a vocal tract, different vocoid shapes can result in different F-pattern outputs with extremely varied Fourier spectra. Except for a few recondite properties like the third moment about the center of area of the intensity vs frequency spectrum, the audible attributes of vocoid frequency spectra are identified in the literature (section 2.2.1) with some of the following parameters (symbolized as in Fant, 1970, p. 241) or with certain sums, differences, or ratios of these parameters:

- F_0 = frequency of the voice source fundamental in Hz;
- L = level of the overall transfer function in dB;
- L_n = level of formant n in dB, the level at $f = F_n$;
- F_n = frequency of formant number n in Hz;
- FN = formant number n without reference to its dimensions;
- B_n = bandwidth of formant n in Hz.

From the acoustic theory of speech production (Fant, 1970), it can be shown that at all frequencies, f , the transfer function $H(f)$ is calculable from equation (1.1a).

$$(1.1a) \quad \omega_n = 2\pi F_n \quad \sigma_n = \pi B_n \quad j\omega = 2\pi f$$

$$s_n, s_n^* = (\sigma_n \pm j\omega_n)$$

$$H(j\omega) = \prod_{n=1}^{\infty} \frac{s_n \cdot s_n^*}{(j\omega - s_n)(j\omega - s_n^*)}$$

The magnitude transfer function can be calculated by equation (1.1b).

$$(1.1b) \quad \begin{aligned} L(f) &= 20 \log_{10} |H(2\pi jf)| \\ \text{in dB} & \\ &= \sum_{n=1}^{\infty} \left[20 \log_{10} \left(\frac{F_n^2 + (\frac{1}{2}B_n)^2}{\sqrt{(f-F_n)^2 + (\frac{1}{2}B_n)^2} \cdot \sqrt{(f+F_n)^2 + (\frac{1}{2}B_n)^2}} \right) \right] \end{aligned}$$

An examination of Equations (1.1) reveals that L and all the L_n 's and hence all their sums, differences, and ratios are derivable from the F-pattern, F_1, F_2, F_3, F_4 , etc. and the associated bandwidths, B_1, B_2, B_3, B_4 , etc. Furthermore, to a first approximation, all the information in the transfer function of a vocoid vocal tract configuration is predictable from the formant frequencies alone, because formant bandwidths are fairly predictable from formant frequencies. The studies of Dunn (1961) and Fujimura and Lindquist (1971) yield the approximate relation given in Equation (1.2).

$$(1.2) \quad \begin{aligned} f &= F_n \\ B_n &\equiv \begin{cases} F_1 < 400 & : 107 - (f/6) \\ 400 < (F_1 \text{ or } F_2) < 1000 & : (f/100) + 36 \\ 1000 < F_2 & : (f/50) + 26 \\ F_3 & : (f/7) - 220 \end{cases} \end{aligned}$$

More complicated and more accurate formulas for predicting formant bandwidths from F-patterns have been developed by Fant (1972, p. 47).

Supra-laryngeal articulatory shifts primarily change the F-pattern, and all acoustic parameters of the transfer function, even the recondite ones, are predictable from the F-pattern. Examples of some analytical relations within vocoid transfer functions include the following:

$$m \neq n$$

- 1) L is a function of the frequency of the formants, but mostly of the lowest formant frequency, F_1 . As F_1 goes up in frequency, L increases. For instance, as F_n doubles, $L(f)$ increases by 12 dB for $f > 2F_n$, all other F_m fixed.
- 2) When two formants get close in frequency, their levels increase considerably, as does the amplitude in the valley between them. For relatively small $(F_m - F_n)$, as $(F_m - F_n)$ is reduced by $1/2$, L_m and L_n are increased by 6 dB and the valley between by about 12 dB.
- 3) The level of a formant of fixed bandwidth increases 6 dB per octave of frequency increase, all other formant frequencies fixed.
- 4) As B_n increases, L_n decreases, with all other factors held constant.
- 5) The quantitative particulars of the above four relations are derivable from equation (1.1b). In order to match measured levels for real vocoids or for the synthetic sounds used in our experiments, a general -6 dB/octave fall-off due to combined source and radiation characteristics must be taken into account.

See Figure 1.2-1 which displays selected source, filter, and output spectrum envelopes that exemplify some of these relations.

Returning to the analogy with the perception of rectangular shape, we might say that as the perception of the shape of figures can be studied apart from a potentially significant signalling feature like color by holding color constant, so the perception of filter functions can be studied separately from features of the glottal source by holding the source constant. Also, as the perception of rectangular shape might be organized by stimulus features that do not allow independent factorial realization, as with height, area, and height/width; so vocoid transfer functions might be heard in terms of some combination of formant frequencies and formant levels, although some dimensions in such a set would be completely predictable from the others.

1.3 Phonetic Influence on Vowel Systems.

After differentiating the phonetic from the formal aspect of phonology, this section focuses on whether or not any one representation of vocoids at one point in the phonetic channel is adequate to explicate the phonetic influences on vowel phonology.

The formal aspect of phonology is evident in distributional rules for phonemic grouping, in debates over the binary vs n-ary nature of features, and in debates about the generative vs interpretive nature of phonology in general. Positions on issues like phonological rule ordering and the abstractness of underlying morphological forms are essentially formal. Such formal principles can be assumed to influence phonemic systems or at least to influence phonological solutions. For instance, most phonologists, including Trubetzkoy (1969, p. 59) and Pike (1947, ch. 10,12), describe a pressure toward formal phonemic symmetry, and proponents of strict binary opposition in phonology see this formal binarity requirement influencing the sequence of sound systems that emerge in children's speech.

There is also a phonetic aspect of phonology which is reflected in the search for an adequate universal phonetic framework, in the common requirement that members of a phoneme be phonetically similar, in the view that historical and sequential sound mutations must be phonetically mediated, or that phonemic norms must maintain their phonetic distinctness above a certain criterion. The extent of agreement among linguists about the influence of the articulatory-acoustic-auditory channel on linguistic form can be measured by the relative isolation of Hjelmslev's (1936, p. 49) viewpoint that "no extra-lingual criteria can be relevant [to phoneme identification], i.e., neither physical, nor physiological, nor psychological criteria." Many phonetic taxonomies - from Bell's 'Visible Speech' to Chomsky and Halle's 'Chapter 7' - have been proposed and applied to explain phonological regularities.

This dissertation reports an investigation into the influence that phonetic properties of vocoids might have on paradigmatic systems of vowels. There already exist numerous a priori phonetic

classifications of vocoid quality. These systems usually classify vocoids at a single stage (articulatory, acoustic, or auditory) of the phonetic channel and assume that such a classification is sufficient to allow the elegant organization of all phonetic facts and to account for all the ways in which phonetic structure can influence phonology. Three approaches to the phonetic taxonomy of vocoids are current:

- 1) some linguists have suggested that the acoustic nature of vocoids is most basic (some not clearly distinguishing "acoustic" from "auditory");
- 2) others have emphasized articulatory structure; and
- 3) a third view is seen in claims that the articulatory and acoustic representations are approximately homomorphic.

The following review of these three approaches to vocoid phonetics leads to the conclusion that nothing less than the elaborated structure of the whole phonetic medium is adequate to organize the various phonetic influences on vowel phonology.

1) Jakobson, Fant, and Halle (1963, p. 12-13) propose that the auditory stage of the phonetic channel is most pertinent for phonology. Many authors since that time have adopted this view of the primacy of auditory perception in the definition of vowel quality and vowel systems, e.g., Fromkin (1968, p. 164) or Liljencrants and Lindblom (1972). Indeed, even speaking in terms of compensatory articulation of a "speech sound" or in terms of articulatory variants of "a vowel", is identifying vowel quality with an acoustic manifestation and its auditory result.

The use of the Cardinal vowel diagram also gives evidence of the essentially auditory 'nature' of vocoids. The Cardinal vowels were framed in an articulatory terminology before radiographic study of articulation was common. When Russel (1928) and Daniel Jones (1932) found on inspection of x-ray pictures of the vocal tract that the highest point of the tongue did not correspond well with the positions on the Cardinal vowel chart, Jones maintained his Cardinal vowel system, which had proved to be of great heuristic value, but changed the definitions of the intermediate Cardinal vowels into auditory terms. It later turned out that the Cardinal vowel chart was closer

in form to a plot of the vowels in an F_1 by F_2 plane. That the Cardinal vowel system with its largely auditory basis has good inter-phonetician reliability has been demonstrated by Ladefoged (1967, p. 141).

The primacy of acoustic phonetic form in lexical borrowing is upheld by Holden, who argues convincingly (1972, p. 111) that phonetic assimilation in borrowing operates through an interpretation of acoustic structure by speakers of the target language.

2) The traditional and currently predominant view is that the articulatory stage of the phonetic channel is the most important one for phonology. Chomsky and Halle (1968, ch. 7) have proposed a 'phonetic framework' for phonology that is completely articulatory and which has had great influence on phonological theory and practice since its appearance. For this reason much current phonological research is framed in exclusively articulatory terms.

Many acoustic phonetic relations also point to the importance of the physical articulatory mechanism in determining vowel quality. For instance, Fant (1966) found that the scaling factor for relating male and female formant frequencies was not constant for all vowels, but that it was greater for some and less for others. This scaling problem had no simple solution in the acoustic domain, but it was shown to be soluble in the articulatory domain. It seems as if American and Swedish women do not construct an auditory vowel system and then shift it up in frequency to a more comfortable frequency range; instead, Fant demonstrated, their vowels sit as if they construct an articulatory vowel system which is shifted up in frequency by a smaller overall vocal tract size and is distorted in frequency relative to the average adult male system by differences in the average proportions of women's vocal tracts. However, Kahn's (1975) data from Arabic speakers suggest that the effect of different anatomical proportions will not account for all sex differences between vowel systems either.

Another phonetic effect that cannot be adequately handled in acoustic terms is vowel reduction. Lindblom (1963, p. 93) concludes his study of the acoustic patterns of vowel reduction by saying,

"once more, acoustic phonetic experimentation has demonstrated that neat and simple organization of data cannot be achieved unless we turn our attention to the preceding stages" (i.e., articulation). In a similar vein, Öhman (1967) devised a single formula in articulation that accounted for a complex set of disparate acoustic patterns found in VCV coarticulation.

Inspection of the formant movement in the completely vocoid English words "we" /ui/, "you" /iu/, "oy" /ɔi/, and "yaw" /iɔ/ shows that vocoids do not always move from one target to another by a simple shortest distance route in 'formant space.' (See Figure 4.1-6,7.) However, a shortest distance route between two vocoid articulatory targets (even in a fairly simple vocal tract model) will give the correct acoustic path. These formant movements will be discussed in detail in section 4.1.2.

3) The third view of the locus of vocoid quality is that for many purposes there is a rough homomorphism between important acoustic parameters and important articulatory parameters. Typically, the acoustic parameters are the first three formant frequencies (or some combinations of them) and the articulatory parameters are two physical dimensions of tongue body position, as seen in midsagittal section, and lip rounding. For instance, Massaro (1975, p. 88) confidently asserts that the first two "formants follow directly from tongue height and tongue advancement." Figure 1.3-1a, from Fant (1970, fig. 2.3-5), is based on x-ray measurements and acoustic output from a real person. Lindblom and Sundberg (1969, fig. II-A-10) give data from an articulatory model which indicate that jaw movement is nearly perfectly reflected in F_1 movement and a well-defined tongue deformation parameter affects F_2 almost exclusively (see fig. 1.3-1b). Lip rounding and spreading are well correlated with the respective lowering and raising of all the lowest three formant frequencies in a wide range of circumstances, as schematized in Figure 1.3-1c from Ladefoged (1971, fig. 12). However, if any of the reasons put forward above for the superiority of either stage over the other is convincing, then the structures at the two stages cannot be equivalent. Approximate

correspondences between articulatory and acoustic dimensions are often workable for constellations of static vocoids, but, as mentioned above, they are strained for the acoustic results of dynamic vocoid articulation (see section 4.1.2) and many seeming correspondences disappear when vocoids are placed in a sequential articulatory context including contoids.

In summary, even though the acoustic form of vocoids is a function of their articulatory formation, some phonological processes that are 'intuitive' at one stage are not intuitive at another stage of the phonetic channel. Shifts in vowel quality as a word is borrowed from one language into another may be most clearly seen in auditory terms, and vowel-consonant interactions may only be explicable with reference to articulation. Examples of sound changes that involve various stages of the phonetic channel are discussed by Jonasson (1971) who concludes that "all links in the speech communication chain should contribute to phonological structure, e.g., articulation favors and disfavors certain segments, while perception involves other preferences and constraints (p. 40)." A similar viewpoint is presented by Catford (1974) and by Fudge (1973), and will be adopted here.

As a vocoid is transmitted along the phonetic channel, it changes form, and the structure among its parametric descriptors at different stages may have different implications for phonology. Examples of phonological processes which reflect structure within different phonetic stages can be given. From this, it seems that a phonetic theory should be explicit at each stage and should represent the correspondences among the stages, at least from preceding to later stages (speaker to hearer).

1.4 Project III Explained.

Project III is the application of a general metalinguistic principle to phonetic reality in order to predict (and so explain) phonological facts. The metalinguistic principle is that languages favor segments that are easy to say, easy to hear, and easy to learn; the 'Principle of Least Effort.' The model of phonetic reality is

the product of Projects I and II: a function $G(H(\vec{A})) = \vec{\Psi}$ and the two distance functions $d(v_i, v_j)$. The phonological fact to be predicted is that, given an n -vowel system, some inventories of vocoids are more common (natural, unmarked) than others. In keeping with section 1.3, only well defined procedures that implement constraints from the whole phonetic channel will be investigated. Two hypotheses that formalize the principle of least effort on the elaborated phonetic channel are current in the phonetic literature. These hypotheses are described in section 1.4.1 and illustrated on a hypothetical many-to-one phonetic function in one dimension in section 1.4.2. Section 1.4.3 reviews the evidence originally used to validate the hypotheses and 1.4.4 indicates the properties required in a model of the phonetic channel in order to adequately test the hypotheses.

Note carefully: A large part of Chapter 6 is concerned with questioning whether or not any procedure of the type applied in Project III can satisfactorily predict the shape of vowel quality systems.

1.4.1 One theory for explaining vowel quality systems posits a "Quantal Nature of Speech" (QNS) and has been formulated and tested for consonants and vowels by K. N. Stevens (1972). In another theory, B. E. F. Lindblom (1972) hypothesizes that paradigmatic segment inventories should exhibit maximum perceptual contrast at minimum effort with simple acquisition (Maxi). Lindblom's procedure is designed for complete phonetic inventories, but Lindblom has only illustrated it on vocoid production for vowel systems.

Lindblom's procedure first operates in the auditory stage and then in the articulatory stage. The first step is to locate, for a given number n , the n most auditorily distinct possible vocoids; this is 'easy to hear.' The second step is to identify for each of these n maximally distinct sounds the set of articulatory postures that will produce a close approximation to it. Finally, among each set of auditorily equivalent articulations, Lindblom would choose the articulation that is closest to the relaxed, 'neutral,' position of the vocal tract; this is 'easy to say.'

Stevens' theory of the Quantal Nature of Speech proposes that the phonetic system characteristically has some positions or areas that have relatively stable acoustic-auditory output, and that there are other positions from which a small move in some direction causes a relatively great change in auditory effect. The metalinguistic principle that Stevens uses is that the inventory of acoustic properties that are used for linguistic codes are ones that "can be generated by the articulatory apparatus without requiring precise positioning or maneuvering of the articulatory structures (p. 52)." This is his interpretation of 'easy to say' or 'easy to learn.' In particular, Stevens makes it clear that if a phonetic system could be characterized by a relation of one articulatory parameter $\underline{a}$ to one acoustic parameter $\underline{f}$ as $g(a) = f$, then phonological systems would "seek out" positions of the articulatory parameter where $\left| \frac{dg(a)}{da} \right|$, the absolute value of the derivative of this function, is at its minimum.

1.4.2 This subsection illustrates the operation of Maxi and QNS on $g(a) = f$, a hypothetical one-dimensional analogue to $G(\vec{A}) = \vec{F}$ (as shown in fig. 1.1-1). It will become evident that QNS and Maxi can yield substantially different predictions when operating on the same phonetic structure, $g(a) = f$. Further, it will be shown how changes in the scaling of $\underline{a}$ and $\underline{f}$ (analogous to changes in $H(\vec{F})$ and $\vec{F}^{-1}(\vec{A})$ in fig. 1.1-1) can radically change the inventories predicted by Maxi and QNS, even with the physical function, $g(a) = f$, unchanged.

Figure 1.4-1a presents a hypothetical function of one physical articulatory parameter $\underline{a}$, scaled linearly in cms., into an acoustic parameter $\underline{f}$, scaled linearly in kHz, thus presenting in graphic form an example of a unidimensional analogue to the function G relating the middle two phonetic stages in Figure 1.1-1. Figure 1.4-1 (adapted from Fant, 1970, fig. 1.4-9) shows, in fact, the second formant frequency of a very simple vocal tract model as a function of the distance of a constriction from the glottis. Figures 1.4-1b and c are the same function but plotted on different scales, thus presenting more complete analogues to the whole phonetic channel as in

Figure 1.1-1. In Figure 1.4-1b, the frequency is logarithmic and the articulatory parameter is shown as a hypothetical scale with finer control (or greater perceived distance) toward the center of the scale than at the extremes. Figure 1.4-1c is linear again in frequency but now logarithmic from left to right in a, scaled articulatory distance.

If the QNS rule is applied to the hypothetical phonetic system of Figure 1.4-1 (assuming the scalings represent the relevant proprioceptive measure in articulation and the correct auditory measure of the acoustic result), the four most favored positions are obviously at the points A, B, C, and D. The scalings make no difference, because at those and only those points $\frac{dg(a)}{da} = 0$.

However this method is undefined for some cases. For instance, if there are less than four vowels to be chosen, or there are more than four vowels needed, what happens? What happens if two relatively stable 'quanta' are very close in perceptual result, e.g., if the local maximum at D were lower in frequency, nearer 1.5 kHz? For cases where less than four vowels are needed or there are only three distinct auditory results among four quanta, an obvious solution is to consider $\frac{dg(a)}{da}$ at neighboring points. That is, choose the broader peak or valley, other things being equal. With this rule included in the procedures for determining the favored segments under the QNS hypothesis, the scalings in frequency and articulation are quite important. In Figure 1.4-1a the quanta at A and D are broadest; in 1.4-1b, B and C (or D?) are broadest; and in 1.4-1c, A and B are broadest. For a monotonic relation $g(a) = f$, QNS predictions can be completely scale-dependent.

Lindblom's Maxi hypothesis would operate on the phonetic system shown in Figure 1.4-1 in the following way: 1) for a given number of vowels, it would locate the most distinct possible positions along the f axis and 2) these acoustic values would be realized by articulatory positions along the a axis that are as close as possible to the neutral position, N, on that axis. In Maxi, three vowels would locate themselves in this phonetic system by settling at maximally distinct positions along the f axis; one at 2.2 kHz (the highest

possible frequency), another at .95 kHz (the lowest possible frequency), and one at a point auditorily equidistant from both extremes. The positions of the intermediate vowel(s) are determined by the auditory scaling on a. Thus, the third vowel falls at 1.6 kHz in linear frequency and at about 1.4 kHz in log frequency. Other scales could produce much greater discrepancies. The second part of Lindblom's operation is to find the easiest articulatory positions which will produce these maximally distinct sounds. The ease with which a sound can be made is taken to be its distance from the 'neutral configuration' of the vocal tract. In Figure 1.4-1, the neutral position, N, has been arbitrarily placed at the mid-point along the a axis, although the scale for effective articulatory distance and the neutral position are conceptually independent.

As can be seen in Figure 1.4-1, regardless of the scales, two sounds will always be located at points B and C because these are the only places where the system reaches its extreme high and low frequency output. However, the third sound of three (or the third, fourth, and fifth among five) can be produced at perhaps four different articulatory positions. In Figures 1.4-1a and b, the intermediate sounds will be realized according to Lindblom's theory along the slope from point B to point C, because they can all be produced there and those are the positions closest to N, the neutral position in this model. In Figure 1.4-1c, an intermediate sound would be produced by an articulatory position between points A and B, because this is where 1.6 kHz can be produced closest to N. For five sounds to be located in this phonetic system as scaled in 1.4-1c, Maxi puts them at about .95, 1.3, 1.6, 1.9, and 2.2 kHz with the highest and the lowest at points B and C, 1.6 and 1.9 to the left of B, and 1.3 just to the left of C.

It can be seen that, as with the QNS rule, Lindblom's Maxi procedure is sensitive to changes in scale, even with a fixed physical model of the acoustic effect of articulation, $g(a) = f$. Imagination could devise many more examples of the divergence both within and between the two optimization procedures, just using differently scaled unidimensional functions. But at no stage in the phonetic channel can

$d(v_i, v_j)$ be expected to be a unidimensional interval. As outlined in section 1.1, the phonetic channel is a vector field - it maps from many dimensions into many dimensions. The problem in Projects I and II is not just scaling a given physical dimension, but choosing, combining, transforming, and scaling some several dimensions, and one can imagine how great a difference these choices could make in the predictions from Maxi and QNS.

The next section, 1.4.3, examines the dimensions and scales used by Lindblom and by Stevens.

1.4.3 Original Validating Illustrations. Lindblom's own validation of his Maxi principle is described in Lindblom (1972) and in Liljencrants and Lindblom (1972). Lindblom assumed that a roughly triangular region in the F_1 by F_2 plane is the acoustic range that vocoid articulation can produce and his auditory distance function $d_\psi(v_i, v_j)$ was:

$$d_\psi(v_i, v_j) = \left\{ \left[\text{mel}(F_{1i}) - \text{mel}(F_{1j}) \right]^2 + k \left[\text{mel}(F_{2i}) - \text{mel}(F_{2j}) \right]^2 \right\}^{1/2}$$

with $k \approx \sec \pi/10 = 1.05$

For optimum contrast, the aggregate inverse auditory distance of a system of n vowels was minimized as follows:

$$\text{Minimize } E, \quad E = \sum_{i=2}^n \sum_{j=1}^{i-1} \frac{1}{(d_\psi(v_i, v_j))^2}$$

The distance is just Euclidean distance in (F_1, F_2) with the only concession to hearing being a mel scaling of the formant frequencies. The K factor was used to correct for the influence of F_3 ; however, multiplying F_2 differences by a constant does not yield the "effective second formant," F_2' , as Lindblom curiously suggests. The definition of F_2' is discussed below in section 2.2.2.

Lindblom related articulation to formant space through a sophisticated vocal tract model described in Lindblom and Sundberg (1969, 1971), and he derived sets of articulatory positions that produced [i], [a], and [u] formant patterns. Sets of (jaw opening, tongue shape) pairs that produce [i] or [a] in Lindblom's model are shown

in Figure 1.4-2. Defining "optimal position as that entailing ... minimum displacement of lips and tongue from their neutral positions." (1972, p. 79) Lindblom concluded that optimization selects a close jaw for [i] and an open jaw for [a]. Notice that Lindblom's identification of effort with tongue displacement yields an optimal articulation at the lowest point on the 'constant-F-pattern' lines in Figure 1.4-2. Identification of effort with jaw displacement from neutral would select an optimum at the extreme left of the line, if the neutral jaw position was assumed to be near closed. Perhaps a measure of effort that combined both jaw and tongue deformation from 'rest' position would be more adequate.

In Stevens' presentation of the "The Quantal Nature of Speech" (1972), he demonstrates the QNS principle in a manner converse to our illustration in section 1.4.2, showing that if a segment like [a] is commonly found in vowel systems, then it will involve a local minimum in the gradient of $G(\vec{A}) = \vec{F}$. Stevens' illustrative function from $\vec{A} = (b, d)$ into $\vec{F} = (F_1, F_2)$ is reproduced in Figure 1.4-3. Stevens asserts that for the best [a] in the model, only when $\underline{d}$ changes alot does the auditory result change appreciably (i.e., when $\underline{d} = 8.5$ cm, $\frac{\delta G(\vec{A})}{\delta d}$ is small).

The acoustic characteristics of two-section resonators like the one shown in Figure 1.4-3a are discussed in Fant (1970, section 1.41) and from Fant's calculations a problem with Stevens' model becomes evident: the input-output relations, $H(\vec{A})$, of two section vocal tract models are not sufficiently similar to the behavior of real vocal tracts. The acoustic pathways resulting from linear movements in that two parameter articulatory space are not human-like; or put another way, the partial derivatives in the model do not correspond well to any directional derivative in a correct model of the vocal tract.

A second problem with Stevens' original validation of the QNS rule for vowels is imprecision in specifying auditory distance. It seems that Stevens assumed, as did Lindblom, that auditory distance for vocoids is some linear combination of formant frequency differences. Notice that such a distance formula would roughly equilibrate the two

hypothetical gradient minima in Figure 1.4-4a. However, if L_2 is considered an important potential auditory attribute (see Flanagan, 1957, 1965), then the relative "peakiness" of the two hypothetical minima are quite different, as in Figure 1.4-4b.

1.4.4 Conclusion: Adequate Specifications. As is probably obvious to both Lindblom and Stevens, in order to test the Maxi and QNS hypotheses, we need more precise specifications of perceived distance in formant space and articulatory space than were used by these authors. Particularly for QNS, a more sophisticated vocal tract model for the function $H(\vec{A}) = \vec{F}$ must be used.

When Maxi is carried out, the first step is to maximize the sum of inter-vocoid distances within a region of formant space that is shaped according to the possibilities of human production. In order to add distances, the distances must be unique up to a linear transformation. For the last step in Maxi, when from among a set of articulatory positions that produce reasonable approximations to a selected F-pattern, the closest to neutral is selected, only an ordinal measure of articulatory distance is required. But since both the perception and articulation of vocoids are multidimensional, whether the composite distances are required to be interval or ordinal, the dimensions must be scaled to a common unit.

Besides a realistic vocal tract model, QNS also requires an interval level auditory distance and proprioceptive articulatory distance in order to construct the gradient partials $\frac{\delta H(G(F(\vec{P})))}{\delta p_i}$.

The remaining principal requirement in testing these hypotheses is that the phonetic models on which they are applied be as phonetic (non-linguistic) as possible. Chapter 2 reviews some relevant previous research, focussing on freedom from linguistic contamination and on the precision requirements of Maxi and QNS. Chapter 2 justifies the approaches to Projects I and II discussed in Chapters 3 and 4.

CHAPTER TWO

METHODOLOGICAL REVIEW

Chapter 2 will formulate Projects I and II in more detail. The projects will be introduced both in light of previous related research and as directed toward the requirements of Project III. Since the ultimate product of Projects I and II is a compound function representing the correspondence between articulatory distance and auditory distance for a set of sounds, and since these distances in the phonetic channel should be phonetic as possible in order that Project III not be circular, methods for obtaining and modeling human data with minimal language-dependence will be emphasized. Sections 2.1 and 2.2 develop Project I from a review of current knowledge about the auditory structure of static voiced vocoids.

The usage of "vowel" and "vocoid" in Chapter 2 generally followed the usage of authors whose work is under discussion. The term "vocoid" has not been commonly used in the speech perception literature and the distinction between a physical and a linguistic perspective has often been ignored or has been out of focus.

Section 2.3 introduces measurement theory and section 2.4 develops Project II from a consideration of previous research in relation to the need for a phonetic model of articulation.

2.1 Language Independence in Perception

For non-circular application, meta-linguistic rules must apply to inputs that are as language independent as possible. Only in this way can phonetic structure 'explain' phonological structure. A phonologically motivated phonetic structure, e.g., as outlined by Jakobson, Fant, and Halle (1963), cannot turn around and claim to explain its own basis. In attempting to clarify what a phonetic, i.e., language-independent, study of vocoid perception would mean, a terminological problem is encountered. In the speech perception literature, "phonetic" is often taken to mean 'speech-related.' In this way, Pisoni (1973) and others differentiate "auditory" from "phonetic" perception approximately as we would differentiate

"phonetic" from "phonological" perception. Because "auditory" is unambiguous, it will refer here to phonetic (i.e., non-speech mode) aural perception.

Methods for obtaining auditory information. Whether one gets purely auditory or more phonologically contaminated information from a listener is largely dependent on the choice of stimuli and procedure. In general, experimental subjects can handle vocoids as auditory sensations much more readily than they can handle contoids as auditory phenomena. It is difficult for listeners to react to contoid-like sounds as anything but linguistic events.

When listening to vocoids, experimental subjects respond in a more auditory mode when the stimuli are longer in duration (perhaps 300 msec, rather than 50 msec), when stimuli are more steady state (no formant frequency changes and small source spectrum changes) and in general when the stimuli are less 'realistic.' Procedures for minimizing possible phonological effects include manipulations of stimulus presentation format and response format. From the results of Barclay (1972) and Pisoni (1973a,b), it seems that stimulus presentations that minimize the role of memory also minimize the language relatedness of subjects' responses; as if speech-like stimuli are coded in memory in a phonological form (Wickelgren, 1965). From this it may be supposed that, for instance, AB presentations are more conducive to auditory judgments than are ABX presentations. Shorter interstimulus intervals may also allow more purely auditory responding than longer interstimulus intervals.

The last, most obvious, and perhaps most important procedural variable is the response type. "Yes-no," "same-different," or numbers are less likely to introduce phonological contamination than are linguistic or orthographic identifications like "heed-who'd" or "i-u." For this reason, confusion matrices usually present more language-influenced data than would be found in differential discrimination data, for instance. Notice also that a non-linguistic response format can more easily accomodate listeners from different linguistic

backgrounds. The more important auditory qualities of vocoids should be extractable from experimental data produced by speakers of any language. For an extensive comparison of several presentation and response formats in vocoid perception, see Hanson (1967).

Keeping the above in mind, we can examine some current hypotheses about the auditory quality of vocoids, while focussing on the purposes of Project I: identifying and scaling a set of auditory dimensions and combining these into an auditory distance.

2.2 Auditory Quality of Vocoid Transfer Functions

Current theory regarding the auditory quality of vocoids is not very diverse. Most investigators accept the idea that the auditory impression created by a vocoid is composed from the effects of a few auditory dimensions and that each of these dimensions is a simple function of one or more formant frequencies. Some specific hypotheses devised or accepted by phoneticians and phonologists will be reviewed; first in relation to the different kinds of evidence that support the hypotheses (section 2.2.1), and second in relation to the different empirical (predictive) content of the hypotheses (section 2.2.2).

2.2.1 Current Theory. In this subsection theories about the acoustic basis of vocoid perception will be organized according to the kind of evidence that supports them. The first group of theories is based on evidence either from phonological properties of vowels or from phonetic properties of phonemic norms for vowels. The second group of theories are experimentally based, although the experimental approaches themselves are diverse.

1) Phonological basis. The most usual way that phonetician/phonologists express their convictions about the auditory structure of vocoids is through their method of presenting formant data. If anything is said about an acoustic vowel chart used to display a set of vowel norms, it is usually that the scales and dimensions were chosen to best approximate the investigator's sense about the real relationships in the language. Peterson (1961, p. 13-14) argues that "the choice of

scales for F_2 vs F_1 formant plots must be based primarily upon the resulting distribution (or spacing) achieved for the various vowels it is desired to represent." Peterson found mel or log scales satisfactory, except that these scales do not allow a good first order correspondence for vowels from different sized speakers. For this reason, Peterson usually used a dual-linear plot. A dual-linear formant chart has both F_1 and F_2 plotted linearly in Hz, but F_1 is scaled twice as large as F_2 . In a formant chart that uses a dual-linear scale, a 100 Hz interval along the F_1 axis is equal to a 200 Hz interval along the F_2 axis. This usage can be extended, so that "tri-linear" means a scale in which a 100 Hz difference in F_1 equals a 200 Hz difference in F_2 which equals a 300 Hz difference in F_3 . Peterson, however, occasionally used a $(\log F_1)$ vs $(\log F_n/F_1)$ display.

Labov, Yaeger, and Steiner (1972) also map vowels onto a dual-linear display, remarking that "this display gives proportions which closely resemble the distribution we have been accustomed to in impressionistic transcription oriented to the cardinal vowels" (p. 31). Carlson, Granstrom, and Fant (1970, p. 35) plot F_1 against a complicated transform of the F-pattern, F_2' (see equation 2a). They comment that "a mel scale representation mel F_1 versus mel F_2' of the [Swedish] vowels brings out certain regularities and equidistant spacings in the vowel system."

Ladefoged has shown an interesting development from plotting vowels in mel F_1 vs mel F_2 (1964, ch. 9) to plotting vocoids in mel F_1 vs mel F_2' (1971, fig. 12), to defending (1975, ch. 8) the auditory and phonological reasons to plot vowels in $(\text{mel } F_1)$ vs $(\text{mel } F_2 - \text{mel } F_1)$. Fairbanks and Grubb (1961) plot the English vowels in $\log F_1$ vs $\log F_2$.

Jakobson, Fant, and Halle (1963) offered a set of auditory dimensions of speech sounds that were primarily based on a complicated body of phonological evidence. Two notable aspects of the JFH auditory dimensions are their applicability to both contoids and vocoids and their suggestive names which reflect an unrealized intention that these dimensions would later be refined into more exact and

general psychophysical transforms of spectrum envelop shapes. As they stand, these features (dimensions) are still identified with their initial, tentative definitions (1963, p. 26-39): compactness $\approx F_1$; gravity $\approx F_2$; flat-sharp $\approx F_2 + F_3 + F_4$; tenseness $\approx |F_1 - 500| + |F_2 - 1500| + |F_3 - 2500|$. Their nasality feature does not apply here. Stevens and House (1972) maintain JFH's emphasis on and ambivalent definitions of compactness, gravity and tenseness; while adding fundamental frequency and duration as important potentially independent auditory qualities of syllabic nuclei.

II) Experimental bases. Experimental studies of the auditory qualities of vocoids have taken many forms. There have been about four principal experimental approaches: (a) synthesis for good phoneme replication; (b) synesthetic scaling; (c) general multidimensional scaling; and (d) fine-grained or psychoacoustic testing.

a) Attempts to synthesize identifiable vowel-like sounds go back over 100 years, but it is reasonable to start with the work of Miller (1953) and Delattre et al. (1952) who adjusted the formant frequencies and formant levels of two-formant synthetic sounds to find the best matches to Cardinal, English, and French vowels. Delattre's matches to the Cardinal vowels are given in Figure 2.2-1. The most important result from these studies was that the best two-formant match to the front vowels (vocoids with high F_2) was attained by positioning a lower formant at exactly the natural F_1 , but positioning a higher formant at some point between F_2 and F_3 . The authors of both studies concluded that the ear tends to weigh the position of the second and third formant together in the front series and receives from them an overall quality roughly equivalent to that which would be produced by a single intermediate formant. This single intermediate formant that will best match the auditory effect of F_2 and F_3 is usually called the "effective second formant" or F_2' , with frequency F_2' . Another result of interest is Miller's conclusion (p. 121) that "the amplitude relations of the formants are an important factor in the evaluation of steady vowel sounds.... An

increase in the amplitude of a formant (in particular F2 or F3) is interpreted to a certain extent as an increase in the frequency of the formant."

The most complete research on the mechanical and phonological processes relating to F₂' is reported by Carlson, Granstrom, and Fant (1970), who conclude that "a more detailed representation of a vowel spectrum than by two formants [F₁ and F₂'] is probably of greater importance for perceived timbre than for the identity" (p. 34). Their mean matched F₂' for the Swedish vowels are reproduced here as Figure 2.2-2. A formula (from Fant, 1973, p. 52) relating F₂' to (F₁, F₂, F₃) is reproduced below as equation (2a).¹

$$(2a) \quad F_2' = F_2 + \frac{(F_3 - F_2)(F_2 - F_1)}{2(F_3 - F_1)}$$

On the basis of native speaker reaction to synthetic syllables in Arabic, English, and Russian, Paddock (1970) argued that the main acoustic determinants of vocoid quality (and the most important phonological features of vowels) are mel F₁, mel F₂', and (mel F₂' - mel F₁). Paddock noted that all F-patterns, (F₁, F₂'), fall in the same plane in his three-dimensional auditory vowel space, and claimed (p. 145) that "a two-dimensional diagram can be a perfectly adequate representation of the 'space' in which [vocoids] are heard."

b) Synesthetic tests offer another experimental approach to the auditory quality of vocoids. The possible importance of synesthetic correspondences between vocoids and colors was first brought to the attention of the general linguistic community by Jakobson in Kindersprache (1968, p. 73-84) where he maintained that the perception of speech sounds (and especially of vowels) is similar to color

¹The present author would suggest that Fant's F₂' matching results are better fitted by the following formula:

$$\begin{array}{ll} \max(L_2, L_3, L_4) = L_a & \text{For some increasing function } \phi(L_j) \\ \text{mid}(L_2, L_3, L_4) = L_b & F_2' = F_a + \frac{\phi(L_b)}{\phi(L_a)} \left[\frac{F_b - F_a}{2} \right] \\ \min(L_2, L_3, L_4) = L_c & \end{array}$$

perception. According to Jakobson, the two important perceptual dimensions of vowels are (1) greater or lesser abundance of sound - corresponding to the visual dimension chromatic vs achromatic, and (2) the u-i or dark-light process. The brightness (dark-light) opposition was said to be basic, while chromaticity was accessory.

Several partially analyzed experiments conducted by Fischer-Jørgensen (1968) presented phonetically naive Danish listeners with "normal manifestations" of the 11 Danish long vowels. Various groups of listeners were asked to match the sounds to either (1) one of a series of achromatic colors (white to grey to black) or (2) one of a set of standard color chips comprising 11 different hues and 10 different shades of brightness and saturation. Other groups of listeners rated these sounds on 7-point scales between adjective pairs like bright-dark, big-small, tight-loose, hard-soft, light-heavy, thin-thick, etc. At publication, Fischer-Jørgensen was not sure how best to analyze her results, but she suggested that the three most important factors in vowel perception could be identified with $(F_2 - F_1)$, F_1 , and F_2 or F_2' , in descending importance. However she felt that "the auditory, motoric, and phonemic contributions to the perceptual dimensions" still needed to be disentangled (p. 94).

c) Recently, the most common method of investigating the auditory quality of vocoids has been to run an experiment that yields an index of similarity between each possible pair of vowels among a given set of vowels and to analyze all the similarities data at once by a computer program that follows one of a family of procedures known as multidimensional scaling (MDS). The product of an MDS procedure is a Euclidean space composed of some number of dimensions in which the stimuli are embedded such that, on the whole, the distances between the stimuli as they are located in the space correspond to the relative auditory dissimilarities found in the data. An important common property of the MDS procedures is that they all produce ordered sets of dimensions and a statistical index of the overall adequacy of the dimensional solution as a summary of the data.

The resulting dimensions are completely derived from the response data and assumptions in the MDS model, e.g., that dimensions must be linearly independent (Euclidean, in particular). The dimensions that come out of an MDS program may correspond to a greater or lesser degree to the experimenter's expectations, but the analysis proceeds as if the stimulus had no previously known structure. Findings from three series of experiments that were analyzed by MDS programs are reviewed below; (1) Hanson's search for Swedish phonological dimensions, (2) a pursuit of purely auditory dimensions by Pols and associates, and (3) the again more phonologically oriented research of Terbeek and Harshman.

(1) Hanson (1967) reports the results of 74 experiments which he conducted over a period of seven years. The experiments include a variety of subjects, stimuli, procedures, and data analyses, although the main experiments used a set of 10 synthetic Swedish long vowels presented to young Swedish adults for magnitude estimation of the dissimilarity of pairs. Of methodological interest, Hanson found no difference in results from children versus young adults, no difference by using ordinal triadic comparison versus paired comparison and magnitude estimation, and he got the same solution for his data from several different MDS programs. Of more interest to us, his substantive conclusion was that, since a fairly constant three-dimensional configuration of his stimuli kept appearing in his MDS analyses, there must be three auditory dimensions in vowel perception; the most important one corresponding to F_2 , the next most important one corresponding to F_1 , and a third one that he claims is "not directly correlated to a known feature or physical variable" (p. 162). Hanson suggests, however, that the third dimension identified by the MDS programs might be interpreted as F_{3rel} , a monotonic scaling of F_3 calculated as follows:

$$F_{3rel} = \frac{(F_3)^2 - (\bar{F}_3)^2}{(\bar{F}_3)^2}$$

(2) Pols and his associates have investigated the physical and perceptual qualities of the Dutch vowels in a series of studies (Pols, van der Kamp and Plomp, 1969; Pols, Tromp, and Plomp, 1973). In their most comprehensive perceptual study, Pols et al. (1969) tried to present their natural vocoid stimuli in such a way that only the information in the supra-glottal transfer function was varied. For this purpose they cut one pitch period from the center of each of 11 CVC Dutch words as spoken by one of the authors and sampled each excised pitch period

in such a way that [they] got the same number of samples (162) for all vowel periods. [At 20,000 samples/sec] this resulted in a fundamental frequency of 123.5 Hz for the vowels.... By repetition of one period a fixed number of times, duration of all signals was made exactly the same (405 msec). Five [preliminary] subjects matched the loudness of all stimulus pairs...[and] by correction of the amplitudes, the loudness levels of all 11 stimuli were made equal. (1969, p. 462)

The stimuli were then presented to Dutch adults in listener controlled continuously available triads and the resulting ordinal data was analyzed by a MDS routine.

It should be pointed out that, although the method of stimulus presentation and data collection used by Pols was nearly optimal, their stimuli are problematic. (1) The method for equalizing fundamental frequency introduces systematic changes in all formant frequencies, F_n , (2) Some over-all amplitude (and presumably loudness) differences are dependent on the transfer function, see section 1.2.2, and (3) Because the glottal source of excitation to the vocal tract characteristically has frequency spectra with various irregularities from one pulse to the next, the repeated presentation of one particular pitch period introduces a fixed but quasi-random frequency domain overlay on top of the vocal tract transfer function.

A physical analysis of the vocoids is reported by Pols et al. in order to approximately simulate the information available at the cochlea. Their cochlear analogue was a bank of eighteen 1/3-octave band pass filters. The physical analysis was accomplished by playing

the stimuli into the bank of 1/3-octave band pass filters and then doing a principal components statistical analysis on the output amplitudes of the filters. A principal component is thus one weighted sum of the outputs of the eighteen filters, or equivalently the output of a peculiarly shaped filter.

The MDS analysis of their perceptual data yielded three relatively important dimensions which "are highly correlated to F_2 , F_1 , and L_2 ," but Pols concludes that there are "only two independent factors, being F_1 and F_2 [because] L_2 is negatively correlated to F_2 " (p. 465). Like Hanson, they say "a description of the third dimension is hard to give." The reader should note that although in the sample of pitch-periods from eleven vowels used by Pols et al., L_2 may be negatively correlated with F_2 , this is not a general (or analytical) relation, as can be shown from Equation 1.1b in Section 1.2.2. In general, ceteris paribus, L_2 is a U-shaped function on F_2 .

Originally, Pols et al. were more inclined toward their "proposed dimensional analysis of spectra based on a 1/3-oct frequency analysis, deliver[ing] three or four well-defined factors [which] are also in good agreement with the results of a perceptual evaluation of the sounds by observers" (1969, p. 466). However, Pols et al. (1973) conclude "that F_1 and F_2 are the most appropriate" parameters for spectral description and that "the optimal plane derived from the 1/3-oct [power levels] is equivalent with the $\text{Log } F_1$ vs $\text{Log } F_2$ plane" (p. 1100).

(3) Terbeek and Harshman (1971, 1972) used Harshman's PARAFAC MDS program to study the perception of a set of twelve vocoids by groups of American, Thai, German, Turkish, and Swedish listeners. The twelve vocoid stimuli were produced by a phonetician and subsequently sampled with eight-bit resolution and low-pass filtered at 3500 Hz. The stimuli were presented to the listeners for triadic comparisons. The first MDS analysis (1971) of the resulting data yielded about three 'unique' dimensions for each listener group, with the first two dimensions interpretable (perhaps as F_1 and

($F_2 - F_1$) and the third MDS dimension corresponding to "rounding" for the German listeners, but uninterpretable for the other listeners.

Terbeek and Harshman (1971) conclude, in part, that

a good deal is known about the acoustic and articulatory structure of vowel sounds, ... if the perceptual dimensions cannot be related to any known characteristics of the stimuli, it is questionable whether the analysis has rendered any useful information at all. (p. 31)

They further suggest (p. 37-8) "that the third factor is an artifact of the linear model used by PARAFAC, and that [possibly the data] can be re-described in terms of two oblique axes."

In a second paper, Terbeek and Harshman (1972) raised more basic issues relating to the perception of vowels: For instance, might not the perceptual dimensions be non-orthogonal and might not their rule for combining into distances be non-Euclidean? By re-analyzing data from Pols and Hanson and data from their own listener groups, Terbeek and Harshman claim "to have found a closer approximation to the 'true' vowel space," (p. 19) by uncovering a correlation between the third (usually uninterpretable) dimension, D3, and the two basic dimensions, D1 and D2. With the origin of the PARAFAC space near the center of the stimulus set, D3 was shown to be well-correlated to peripheralness as measured by squared distance from the origin, $D1^2 + D2^2 = D3$.

Since the value of a vowel on dimension D3 is predictable from a function of its value on D1 and D2, the vowels are necessarily restricted to a curved surface embedded in the three-dimensional space. One might suspect then, that the third dimension is not a third perceptual criteria, but an artifact of the model's restriction to linear dimensions. (p. 19)

Terbeek and Harshman propose that

$$\text{if } D3 = f(D1, D2) = k(D1^2 + D2^2) + c,$$

then the distance between two vocoids, v_i and v_j , should be expressed as:

$$d(v_i, v_j) = \left[(D1_i - D1_j)^2 + (D2_i - D2_j)^2 + [f(D1_i, D2_i) - f(D1_j, D2_j)]^2 \right]^{1/2}$$

The two arguments of this distance may be phonologically reasonable ($\pm$ high, $\pm$ back) and the troublesome third dimension is replaced by a troublesome combination rule. But other problems remain. 1) D3, as peripheralness, is not uninterpretable nor phonetically implausible. JFH's [$\pm$ tense] feature is monotonically related to D3, and Labov, Yaeger, and Steiner's (1972) [$\pm$ peripheral] feature is certainly related and phonologically plausible. 2) Terbeek and Harshman noted that D1 and D2 did not order the vowels in height (F_1) or in advancement ($F_2 - F_1$); D1 and D2 merely allowed a binary division of the vowels (for most speaker groups) into [$\pm$ high] and [$\pm$ back]. If D1 and D2 cannot be verified as continuous (phonetic) parameters, how can D3 be a continuous function of them? 3) D3 is an "artifact" of the linear combination of dimensions assumed in the PARAFAC MDS program, only if D3 is necessarily constrained by D1 and D2. However, correlation between D3 and $f(D1, D2)$ was shown only for the 12 stimuli used in their experiments. Among the stimuli used by Pols et al., L_2 was well correlated with inverse F_2 , but in general F_2 and L_2 are not monotonically related.

d) Finally we consider two experimental investigations into vocoid perception which are more psychoacoustic. The most complete studies of the auditory psychophysics of vocoids have been conducted by Flanagan (1955a,b; 1957a,b; 1965) in order to estimate the precision needed for high quality frequency domain vocoding. Flanagan determined difference limens for F_1 and F_2 , L_2 and L , and F_0 (fundamental frequency). Flanagan also determined difference limens for the amplitudes of selected single harmonic components of vocoids. Flanagan's most important finding was that the difference limen (DL) for F_2 is markedly assymmetric when F_2 is near F_1 or near F_3 in serially synthesized sounds (i.e., stimuli behaving according to Equation 1.1b). Flanagan suggested that although the formant frequency DL is generally about 5%, when

the first and second formants of the synthetic vowel are relatively close, negative deviations of F2 result in comparatively large increases in amplitudes of F1 and F2 and hence a small negative frequency DL. Similarly, [when] the second and third formants are in close proximity, positive deviations of F2 result in comparatively large increases in F2 and F3 amplitudes and hence a small positive frequency DL. These effects obtain, of course, in spite of the fact that over-all intensity levels are maintained constant [in this experiment].
(1955a, p. 616)

After observing a ± 1.5 dB DL for vocoid overall intensity, L, Flanagan (1957a) tested the DL for L_2 in parallel synthesized sounds (which allow L_2 to be varied independently of F-pattern). The DL was found to be about ± 3 dB, while an 8 dB increment in L_2 concomitantly increased L by less than 0.5 dB, thereby demonstrating the perceptual salience of formant level even without a perceptible general level increase. The asymetries in the F_2 DL were explicable, and Flanagan (1957b) seems to have tacitly accepted the conclusion that vocoid transfer functions have three main acoustic "dimensions" (F_1, F_2, F_3) which are perceived in six main auditory "dimensions" ($F_1, F_2, F_3, L_1, L_2, L_3$). Assuming six dimensional transmission in a vocoder for parallel synthesis at the receiving end, Flanagan takes the ranges of F_n and L_n to be

$$\begin{array}{ll} F_1 : 250-800, & L_1 : 5 \\ F_2 : 800-2300, & L_2 : 20 \\ F_3 : 1700-3000, & L_3 : 20 \end{array}$$

with the following "rough estimates" of the maximum precision necessary in quantizing the 'dimensions':

	step size	quantization		step size	quantization
F_1	± 20 Hz	14	L_1	± 1 dB	3
F_2	± 50 Hz	14	L_2	± 2 dB	3
F_3	± 75 Hz	9	L_3	± 5 dB	2

The step sizes are given in Hz, which would indicate a tri-linear auditory frequency scale, but in this (1957b) text, and later,

Flanagan cites formant frequency DL as 3% which, along with his use of logarithmic scales in his pictures of vocoid spectra, indicates an assumed logarithmic auditory frequency scale. One might assume an additive combination rule among the six auditory dimensions and infer a common unit approximately equal in each dimension to the step size. If these are fair inferences, Flanagan has presented the most explicit and quantitative theory of vocoid psychoacoustics available.

A last experimental study tested the adequacy of F_2' in representing the perceptual effects of F_2 and F_3 . Fujimura (1967) presented serially synthesized vocoids that varied in F_2, F_3 to Swedish listeners in an ABX discrimination and an identification test. The stimuli were chosen to sample the acoustic area associated with the 'close front' /i,y,u/ vowels in Swedish. Two variables were manipulated in the experimental design; geometric mean of F_2 and F_3 , and separation of F_2 and F_3 . Eight means of F_2, F_3 in the experiment ranged from 1700 to 3000 Hz, and at each mean four values of separation were presented; F_2 and F_3 being 3, 5, 7, or 9 quarter-tones apart. These 30 stimuli constituted a fairly dense sample of F_2, F_3 near the $F_2 = F_3$ plane in formant space. F_1 was set to 270 Hz and F_4 was set at 3300 Hz and other acoustic parameters were held constant at appropriate settings. "The voice source signal, in particular, was derived from a pulse generator triggered by natural voice signals, so that resulting source signals had a realistically fluctuating pulsation but at the same time a constant flat spectrum envelope" [$F_0 \approx 190$ Hz] (p. 182).

Fujimura notes that this high F_0 results in "a wider separation of components in the harmonic structure of the vowel spectrum [which] may tend to make the quality discrimination less sensitive," concluding that "even with this 'unfavorable situation,' our results show that the subjects still were concerned with the local structure of the spectrum envelope" (p. 191).

As an ancillary finding of the ABX tests, Fujimura suggests that proximity of F_2, F_3 (with resulting increases in L_2 and L_3) can be

heard as an increase in mean F_2, F_3 (cf. Miller, 1952, and above).

Fujimura's general conclusion from the identification study was that

any monotonically increasing function in the two variables [F_2 and F_3] would be refuted as a measure for the 'higher effective peak'.... The data at least give strong evidence that we will not find a simple enough formula for deriving one variable (for the high frequency range) that represents the 'effective peak' as the determinant of the vowel quality. Any kind of average of the 'higher formants' would not explain these experimental data. It is highly doubtful if we can find any sensible way of representing the vowel quality in a two dimensional space, like the F_1 vs F_2 diagram as far as the scheme is considered as a language-independent, i.e., 'universal' framework for phonetic description. (1967, p. 188)

2.2.2 Empirical Differences. This subsection will illustrate what differences the choice of various scales, and dimensional representations can make in predictions of ordinal vocoid distances. From the preceding lengthy excursion into the literature, several acoustically based theories of vocoid (and vowel) perception can be enumerated.

Proposed Dimension Sets [optional or secondary dimensions in brackets]

$F_1, F_2, [F_3, F_4]$

$F_1, (F_2 - F_1), [(F_3 - F_1)]$

F_1, F_2'

$F_1, (F_2' - F_1), [F_2']$

$F_1, F_2, (F_2 + F_3 + F_4), (|F_1 - 500| + |F_2 - 1500| + |F_3 - 2500|)$

$F_1, F_2, F_3, L_1, L_2, L_3$

The rules for combining the positions of a vocoid on these dimensions into an overall auditory impression and accordingly calculating an overall distance between pairs of vocoids are usually unspecified in the literature. Delattre, Paddock, Hanson, and Pols clearly assumed a linear combination and Euclidean distance. Terbeek and Harshman

clearly did not. The auditory frequency scales of F_n dimensions are, if specified, typically linear, dual-linear, logarithmic, or mel.

Although a systematic, formal analysis of the general differences among several of the proposed dimensional representations would be ideal, appropriate methods are unavailable. Instead, selected differences in distance predictions will be illustrated below by choosing particular F-patterns and showing how scale differences alone, changes in combination rule alone, and differences in dimensional representation alone can effect the order of inter-vocoid distances.

Scales. Assuming an (F_1, F_2, F_3) representation and a Euclidean distance rule, one can find three different distance orders among three pairs of vocoids by using three common formant frequency scales; dual-linear, logarithmic, and mel. Consider the following six F-patterns which define six vocoids in an (F_1, F_2, F_3) representation:

a: (250,1500,2500)	c: (550,1500,2500)	e: (500,1500,2500)
b: (350,1500,2500)	d: (700,1500,2500)	f: (500,1800,2500).

The three scales induce the following orders among $d(a,b)$, $d(c,d)$, $d(e,f)$:

dual-linear	$cd = ef > ab$
logarithmic	$ab > cd > ef$
mel	$ef > cd > ab.$

Combination rules. Without settling on any particular formant frequency scales, we might consider four different combination rules for distance

d = Euclidean
d1 = paraboloid-cordal
d2 = city block
d3 = a polynomial

and demonstrate their empirical difference. Assuming a dual-linear $(2F_1, F_2)$ representation of vocoids, we may see how a paraboloid-cordal combination rule like that used by Terbeek and Harshman can predict

different orders of distances from a Euclidean rule. In a $2F_1, F_2$ representation, a paraboloid-cordal distance, d_l , might be as follows:

given

$$D3_i = 0.0125((F_{1i} - 500)^2 + (F_{2i} - 1500)^2)$$

$$d_l(j,k) = [(2F_{1j} - 2F_{1k})^2 + (F_{2j} - F_{2k})^2 + (D3_j - D3_k)^2]^{1/2}$$

With $d_l(j,k)$ as above, and Euclidean $d(j,k)$, four vocoids can be easily found, for instance

$$\begin{array}{ll} a: (500, 1100) & c: (300, 1300) \\ b: (500, 1500) & e: (300, 1700), \end{array}$$

for which $d(a,b) = d(c,e)$ but $d_l(a,b) > d_l(c,e)$.

Consider another two combination rules for auditory distances among vocoids in an F_1, F_2 representation with linear scales:

city block distance, d_2

$$d_2(j,k) = |F_{1j} - F_{1k}| + |F_{2j} - F_{2k}|$$

polynomial distance, d_3

$$d_3(j,k) = \frac{F_{1j} + F_{1k}}{200} [|F_{1j} - F_{1k}| + |F_{2j} - F_{2k}|].$$

Given four vocoids

$$\begin{array}{ll} a: (300, 1400) & c: (500, 1200) \\ b: (400, 1600) & e: (600, 1400) \end{array}$$

calculations produce $d_2(a,b) = d_2(c,e)$, however $d_3(a,b) > d_3(c,e)$.

Dimensional representations. More than scales or combination rules for distance, most of the published discussion has been concerned with identifying the dimensions of vocoid quality. How many auditory dimensions comprise vocoid quality and what is their acoustic basis? Popular sets of dimensions were listed above. The number of dimensions in a set ranges from two to six, and all the dimensions are analytically derivable from the F-pattern (F_1, F_2, F_3, F_4), as outlined in Section 1.2.2. Even Pols and associates finally accepted the

superiority of F_1, F_2 as a model of vocoid perception after seeking a solution in their physiologically oriented filter bank model. In principle, even the filter bank output was derivable from the F-pattern.

It may seem strange that two proposed dimensional representations of vocoids like (F_1, F_2) and $(F_1, F_2 - F_1)$, which contain exactly the same information, should constitute different theories of vocoid perception. It is only necessary that they do entail different empirical claims under the assumption of fixed or constrained choices of scales and combination rule. Notice that 'city block' distance on a dual-linear scaling of (F_1, F_2) is perfectly equivalent to 'city block' distance on $(F_1, F_2 - F_1)$ in linear scales. However, if certain frequency scales are applied to both dimension sets and a common distance formula operates on both, then (F_1, F_2) and $(F_1, F_2 - F_1)$ make different empirical predictions. For instance, assuming dual-linear frequency scales and a Euclidean distance for (F_1, F_2) , an F_1 interval of 100 Hz is equal to an F_2 interval of 200 Hz. Under the assumption of a dual-linear scale and Euclidean distance in $(F_1, F_2 - F_1)$, 100 Hz change in F_1 is equal to 224 Hz $[(100^2 + 200^2)^{1/2}]$ change in F_2 .

To illustrate the different empirical consequences of (F_1, F_2, F_3) vs (F_1, F_2') , four F-patterns may be considered; two similar to [i] and [y] in Figure 2.2-2 with similar F_2' values to those indicated, along with two other F-patterns that have low F_2 such that $F_2' \approx F_2$.

[i]	(250, 2000, 3000, 3400)	$F_2' = 3100$
[y]	(250, 2000, 2500, 3400)	$F_2' = 2200$
[o]	(400, 800, 2500, 3200)	$F_2' = 800$
[ø]	(400, 1050, 2500, 3200)	$F_2' = 1050$

Taking Euclidean distance in logarithmic coordinates, the (F_1, F_2') representation predicts $d(o, ø) < d(i, y)$, whereas in (F_1, F_2, F_3) logarithmic Euclidean distance predicts $d(i, y) < d(o, ø)$.

A last example illustrates how the non-monotonic relation of L_2 to F_2 makes for different predictions about vocoid perception as seen through (F_1, F_2) as opposed to (F_1, F_2, L_2) , when scales and combination

rules are constrained. Take the formant frequencies to be scaled logarithmically with a 1 db difference in L_2 equivalent to the F_2 interval between 1500 and 1600 Hz, and consider the three following F-patterns, differing only in F_2 :

a:	(500,1500,2500)	$L_2 = 1$
b:	(500,1800,2500)	$L_2 = 3$
c:	(500,2100,2500)	$L_2 = 8$

Without considering L_2 , $d(a,b) > d(b,c)$, but when L_2 is considered to contribute to auditory distance (scaled 1 db = $d_{F_2}(1500,1600)$), the order is reversed: $d(b,c) < d(a,b)$.

The preceding several examples show that some differences in scales, dimensions, and combination rules result in basic ordinal differences in the empirical content of theories of vocoid psychoacoustics. However, since each dimension in every dimension set is derivable from the formant frequencies, why not treat (F_1, F_2, F_3, F_4) as the psychological dimension set and translate, for instance, Euclidean distance in (F_1, F_2') or in (F_1, F_2, L_2) into a complicated combination rule on (F_1, F_2, F_3, F_4) ? Terbeek and Harshman's abandonment of 'peripheralness' as a dimension in favor of a complicated combination rule on $(D1, D2)$ seems to follow this route.

The following section, 2.3, may help clarify this question and indicate a more general method of identifying empirical differences inherent in some delimitable kinds of numerical theories. A 'measurement theoretic' approach can, for instance, characterize all the proximity orderings that are consistent with classes of combination rules and classes of scales, for a given dimensional representation.

2.3 Measurement Theory and Project I.

Measurement theory (Krantz et al., 1971) is a formal analysis of certain kinds of models that have arisen mainly in psychology and economics. An axiomatic analysis of dimensionality and scaling in relation to ordinal similarity data on multi-dimensional objects has been provided (Beals, Krantz, and Tversky, 1968; Tversky and Krantz,

1970). This analysis suggests, in part, that the definition of dimensionality can be treated as a theoretical claim, entailing certain mathematical forms in a model and certain regularities in proximity data. Usage of the term "psychological dimension" may be restricted to variables in particular kinds of models; thus making "psychological dimension" mean more than just arbitrary 'variable'. The common assumption, embodied in MDS programs, that psychological dimensions combine according to the Euclidean rule, reflects the basic feeling that the order of proximity of pairs of stimuli that differ on only one dimension ought to be independent of the fixed common position of the stimuli on any other dimensions. However, the formal analogue of empirical 'order-independence' is dimensional 'decomposability'. Since the class of decomposable combination rules is a generalization from Euclidean and linear distances, a reliable real-world violation of order-independence can lead to the rejection of not just one particular combination rule, but a whole class of combination rules on a dimension set.

A limitation of measurement theory analysis is that general methods have not been developed for handling non-decomposable combination rules. However, even limited to decomposable combination rules, axiomatic methods of evaluating multi-dimensional proximity are much more powerful than the kind of hit and miss comparisons presented above in section 2.2.2. A further limitation is that experimental tests must be designed within a given dimensional representation of the objects being studied, and efficient experimental designs for testing formal hypotheses should be organized around the distinct predictions of an alternative hypothesis. Thus, designing critical experiments to test measurement axioms requires a well-understood problem and a previously known factorial structure. As Terbeek and Harshman knew, the acoustic and auditory structure of vocoids is well understood (see sections 1.2.2 and 2.2.1), and vocoids can be synthesized to fit the requirements of some factorial designs.

Project I, as described in Chapter 3, tested how the structure of vocoid audition could be related to the three independent variables

(F_1, F_2, F_3) which were manipulated in a series of listening experiments. Measurement axioms were used to design critical experiments which would indicate whether or not vocoid perception could possibly be represented in psychologically scaled counterparts to F_1 , F_2 , and F_3 . By assuming that Flanagan's dimensions, $F_1, F_2, F_3, L_1, L_2, L_3$, are present in the auditory quality of vocoids, measurement axioms allowed the construction of sets of distances in (F_1, F_2, F_3) -space that should be 'congruent' if F_1 , F_2 , and F_3 combined linearly to form distances, but would be 'incongruent' if formant levels (L_1 , L_2 , or L_3) are salient auditory attributes. Stimulus sets that only varied along a single dimension, e.g., F_2 , could also be constructed in such a way that if, for instance, L_2 is perceptually salient, the perceived order of intervals along F_2 would be 'intransitive.' Such data would disallow any monotonic auditory scale on F_2 without the establishment of another dimension (perhaps L_2).

From among the dimension sets listed at the beginning of section 2.2.2, (F_1, F_2, F_3) was chosen to be tested. F_3 was included because it is independently variable and changes in F_3 are audible as shown by the F_2' matches to /y/ vs /i/ and by the general perceptual salience of 'retroflexion' as heard in F_3 shifts. The effective second formant, F_2' , was not tested because, as Fujimura showed, two vocoids with distinct (F_2, F_3) and the same best F_2' match may have distinct timbres. Since differences in timbre (or auditory quality) not phonemic identity, are required for input to Project III, F_2' would not have been appropriate. Vocoid stimuli must incorporate a full F-pattern in order to reliably simulate the natural values of effective L_2 and effective B_2 as well as effective F_2 .

Tests to identify the possible and impossible combination rules for a dimension set like $(F_1, F_2 - F_1)$ are not within the present scope of measurement methods because $(F_1, F_2 - F_1)$ is not decomposable without fixing F_1 and F_2 scales and a scale of $(F_1 - F_2)$ differences. However, some unidimensional tests that can indicate the impossibility of an additive auditory scale for F_2 in (F_1, F_2, F_3) , would also indicate the same impossibility for $(F_2 - F_1)$ in $(F_1, F_2 - F_1)$, and so

disallow Euclidean and certain other distances in $(F_1, F_2 - F_1)$ as well.

Conclusion. Project III requires a psychological distance function on pairs of F-patterns. The distances should reflect the auditory quality (timbre), as much as possible, to the exclusion of linguistic influences. For this reason AB paired comparisons were used as the presentation format in our listening experiments and the response format was numerical. Listeners were chosen from diverse linguistic backgrounds, and most of the stimuli were selected to form well-spaced grids in (F_1, F_2, F_3) without regard to the phonetic norms for vowel phonemes in any particular language.

In the literature on vocoid perception, the research reported by Pols, Flanagan, and Fujimura focussed most directly on the language-independent auditory basis of 'vowel quality.' This research and common practice both seem to indicate F_1, F_2 (and F_3) as the main determinants of vocoid timbre; the other acoustic properties are usually assumed to be secondary, and a Euclidean formula for inter-vocoid distance is often tacitly accepted. In order to fit or construct appropriate scales on the formant frequencies, the dimensional structure and combination rule for distance must be known. The first task in Project I was then to establish or disestablish (F_1, F_2, F_3) as a decomposable auditory structure. Flanagan's work on the perceptibility of formant levels was taken as a clue about where to look for violations of decomposability and additivity in (F_1, F_2, F_3) .

2.4 Measuring and Modeling the Vocal Tract.

Project II centers around building a model of the human vocal tract that would map articulatory positions into F-patterns; i.e., establish a function $G(\vec{A}) = \vec{F}$ as in Figure 1.1-1. The principal requirements for the vocal tract model in relation to its use in Project III are the following: (1) It should accurately reflect human vocal tract shapes and relate them to correct acoustic output, (2) Shapes should be represented in an appropriate parameter set so that movement in the model results in human-like vocal tract movements

with human-like acoustic results, and (3) The model should be as phonetic (language-independent) as possible, allowing a full range of vocoid vocal tract configurations. After consideration of the available options, language-independence was sacrificed to the first two requirements; correct parameterization and correct acoustic output.

Project II also involved an attempt to identify the function $F(\vec{P}) = \bar{A}$, and estimate the relation of physical distance between articulatory positions and the psychological counterparts of these distances, $d_p(v_j, v_k)$. However, since there is little published research on the perception of articulatory position (LaPointe and Wertz, 1974; and Bishop et al., 1973 are marginally related), the proprioception experiments in section 4.2 will have no methodological introduction here.

The measurement of vocal tract shapes in relation to their acoustic results has been extensively studied, yielding results in three areas: (a) the relationship between vocal tract area functions and resonant frequencies of the vocal tract, (b) the relationship between available measurements of vocal shape and actual vocal tract area functions, and (c) the development of an appropriate parametric description of vocal tract shape.

(a) The acoustical effect of the supra-laryngeal vocal tract on the signal which enters it at the glottis end can be completely specified in the frequency range of interest, if its area function (the cross-sectional area of the tract at every point along its length) is known. Methods for calculating the acoustic transfer function of a tube are well understood (see, for instance, Fant, 1970). In the vocal tract model of Project II, area function to F-pattern conversion was implemented by a separate general purpose sub-routine that was available on the computer facility that we used.

(b) Measuring the shape of the vocal tract during the production of vocoids is difficult. Although attempts have been made to directly measure three dimensional shape (Ladefoged et al., 1971) or indirectly estimate three-dimensional shape (Schroeder, 1967; Mermelstein, 1967),

the usual method of measuring vocal tract shape has been based on lateral radiographs. Measurements from the lateral x-ray view are, in turn, used to estimate the area function of the vocal tract. The area function can then be checked by calculating its transfer function and comparing it to the speech wave produced at the time the x-ray was taken. Measurements of the size of the vocal tract in the mid-sagittal plane have been published by Russel (1928) and Perkell (1969) among others, but the principled derivation of area functions from lateral x-rays was first accomplished by Chiba and Kajiyama in 1941 and has been developed further by Fant (1970), Heinz and Stevens (1965), Sundberg (1969), and Maeda (1972).

The various cross-sectional shapes of the vocal tract are derived from other measures of the vocal tract such as plaster casts, palatograms, frontal x-rays, and so forth. Typically, for each different region of the vocal tract, a basic shape is assumed and parameters that determine the area of that shape at a particular point are estimated from the mid-sagittal width at that point. Maeda's method is slightly different from most others in that it uses information from more than one mid-sagittal measurement to determine the parameters of cross-sectional area at particular points.

(c) Identification of the most useful parameter set for describing articulation is a live issue in phonetics. Whether considered for efficient vocoding or for simplest linguistic descriptions, the manner in which vocal tract shapes are specified may have important consequences. Just as the Fourier spectrum of a voiced vocoid can be described by specifying the amplitudes and phases of the lowest hundred or so harmonics, a vocal tract shape can be described by its cross-sectional area or mid-sagittal width at every half-centimeter along its length. It would be better, however, to select a smaller number of parametric descriptors that could provide sufficient detail and bring into focus the structural possibilities of the system.

Omitting from discussion vocal tract models that primarily deal with consonantal actions or articulator acceleration rates (Öhman,

1967; Coker et al., 1973), there are several kinds of parameter sets for describing static vocoid configurations. These parametric descriptions divide themselves into three groups according to their structural basis: some describe arbitrary configurations; some describe a straight tube with one constriction and a lip section; and some have parameters that represent positions of anatomical parts. The available parameterizations of vocoid shape also fall into two groups according to whether they describe the area function or operate to form shapes in a mid-sagittal section.

Descriptions of vocal tract shape in terms of Fourier coefficients have been reported by Schroeder (1967), Mermelstein (1967), and Liljencrants (1971). These descriptions treat vocal tract shapes as arbitrary curves and fit the curves by a small number of sinusoidal components. For suitably transformed mid-sagittal outlines, Liljencrants shows that two numbers (the magnitude and phase of a single sine wave) are sufficient to adequately match the tongue profiles of most Swedish vowels. Schroeder and Mermelstein have shown that Fourier series can also directly specify area functions. There has been no effort, however, to show that descriptions of vocal tract shape in terms of a Fourier series reflect any of the structural constraints to be found in articulation.

For this reason, we will compare an earlier type of vocal tract model as described by Stevens and House (1955) and Fant (1970, sec. 1.43), to a more recent type as developed by Lindblom and Sundberg (1969; 1971) and Mermelstein (1973). The earlier vocal tract models operated directly on a three dimensional area function and had three parameters which could be identified with place of constriction, degree of constriction, and lip rounding. The early models were straight tubes with fixed total length and fixed maximum cross-sectional area. A single narrowing could be positioned at any point along the length of the tube, the cross-sectional area at the point of constriction could be varied, and the length/area ratio of the lips could be modeled by an independent lip section. The acoustic

output of these models in relation to their parametric input has been thoroughly studied. In the Stevens and House model the static positions in the parameter space that best replicated the F-patterns of the English vowels were related to each other in a way that was "highly suggestive" of traditional vowel diagrams. It becomes clear, however, on close examination, that the parameter set of the early models does not produce correct articulatory dynamics. Since the main difference among [ɔ], [u], and [y] in the early models is the distance of the tongue constriction from the glottis (lip rounding and degree of constriction are about the same), the simplest route from [ɔ] to [y] in such a parameter space passes through an [u] configuration. The [ɔ]-[u]-[y] route is illustrated in Figure 2.4-1. Note that a similar route is the simplest one in Liljencrants' Fourier description of tongue profile shapes.

This problem with the early vocal tract models, wherein they travel from 'low-back' to 'high-front' via 'high-back,' can be corrected by representing the L-shape of the vocal tract in the geometry of the model. This is done in the more recent, and more anatomically oriented vocal tract models developed by Mermelstein (1973) and Lindblom and Sundberg. Both recent models manipulate a mid-sagittal outline of the vocal tract, and both have larger and more anatomical parameter sets. The five parameters of the Lindblom and Sundberg model are lip tension, jaw height, larynx height, and two parameters of tongue shape "which can be said to represent point and degree of constriction" (1971, p. 1168). Mermelstein's model describes static non-nasal vocal tract shapes in nine parameters that fix the absolute or relative position of five anatomical parts in a mid-sagittal view: one parameter for the jaw, two for the hyoid, two for the tongue body, two for the tongue tip, and two for the lips.

The advantage of Mermelstein's model is the manner in which it represents tongue body position. The benefit of a more accurate anatomical framework (particularly the L-shape of the tract) is lost if the tongue moves in an L-shaped "semipolar" coordinate system. Simple parametric adjustment in the quasi-semipolar system of the

Lindblom and Sundberg model (1971, figs. 4 and 13) will produce nearly the same [u] route from [ɔ] to [y] as found in the earlier straight tube models. For this reason, Project II used a model which represented tongue body position with spacial coordinates that were independent of the orientation of the vocal tract. Among the available models, a system like Mermelstein's in which tongue position is specified in coordinates relative to some fixed structure like the maxilla is the only kind that predicts the fairly straight tongue movements found by Kent (1972). Kent's Figure 7 indicates that points on the tongue surface move from [a] to [ɪ] in nearly straight lines when viewed in a lateral cineradiograph; the figure shows no tendency toward tongue trajectories conforming to the L-shape of the vocal tract.

Conclusion. A vocal tract model that specifies tongue body position within non-L-shaped coordinates in an L-shaped tract is the only kind that will satisfy the second requirement suggested for vocal tract models; i.e., the parameters of position should be chosen such that simple movement in the model is human-like. However, the first and second requirements are somewhat in conflict with the third requirement. Although the recent models do accurately portray human vocal tract shapes and can be tested for human-like dynamic output, the early models and the Fourier specification of shape are much more language-independent. One problem is that the early vocal tract models had an overly restricted set of possible shapes and the Fourier descriptions have unrestricted possibilities. A more important difficulty is that these parameterizations do not produce the correct routes between vocoid targets.

On the basis of these considerations, Chapter 4 describes a rather language-specific model of articulation. The model manipulates a tongue-body circle with fixed radius by specifying the position of its center in Cartesian coordinates in the mid-sagittal plane. The model is thus a simplified version of the Coker-Mermelstein type, although it is arranged to fit with Maeda's method for conversion of mid-sagittal outlines into area functions. The decision was in favor of

static and possible dynamic accuracy within a limited range, as against an extended range of configurations without the possibility of checking accuracy.

CHAPTER THREE

AUDITORY TESTS

Chapter 3 describes a series of psychophysical listening tests which attempt to clarify the relationship between the acoustic specification of vocoids and the auditory qualities of vocoids as revealed by judgments of relative similarity. First, the acoustic range of the vocoids used in the tests is justified (sec. 3.1). Second, a theoretical approach for testing a class of formal representations of the auditory quality of those vocoids is reviewed (sec. 3.2). Third the general method used in the experimental listening tests is outlined (sec. 3.3), and fourth, the specific experimental tests are described and results given (sec. 3.4-3.7). A last section (3.8) summarizes the experimental results in relation to the axiomatic foundation of various algebraic representations and derives an expression for representing auditory distance between pairs of acoustically defined vocoids. The general reader may want to skip sections 3.4 through 3.7.

3.1 The Acoustic Range of Possible Vocoids

The size and shape of the 'formant space' available for use in languages is well known in general but has not been studied in detail. The ranges of formant frequencies that are heard in speech are quite large if all speakers are measured, i.e., adult men and women, and children. It was decided to conduct the large-scale listening tests on a formant space large enough to include most of the possible F-patterns that could be produced by one adult male vocal tract: F_1 (175 Hz - 700 Hz); F_2 (800 Hz - 2500 Hz); F_3 (1800 Hz - 3000 Hz). These values were chosen in order to enclose as much as possible of the appropriate formant data from Peterson and Barney (1952) and Fant (1970, p. 241, 1973) while minimizing the overlap between the frequency ranges of adjacent formants. The frequency ranges given above yield a formant space that includes a portion of the $F_2 = F_3$ plane, but does not quite reach $F_1 = F_2$. This formant space has rectilinear boundaries except where it intersects the $F_2 - F_3$ plane.

It is assumed that the formant space is 'full' or 'dense', having no holes or unrealizable interior regions (see Paget, 1930, p. 45). Note that this formant space extends toward but does not reach a satisfactory American [L].

3.2 Theory

The experimental auditory tests to be discussed in sections 3.4-3.8 are constructed following an approach to psychological measurement developed by Tversky and Krantz (1970; see also Krantz and Tversky (1971, 1975)). This measurement theoretical approach is based on a network of formal relationships between two groups of statements about vectors (in our case F-patterns); in one group the statements are elemental and ordinal, and in the other group the statements are general and numerical. For example, (a) is an elemental ordinal statement of one-dimensional context independence;

for any two first formant frequencies p, q , and
any four second formant frequencies r, s, t, u :

- (a) if for $F_1 = p$, $d(r,s) > d(t,u)$
then for $F_1 = q$, $d(r,s) > d(t,u)$

and expression (3a) below is a general numerical statement. In explicating the logical connections between these two kinds of statements, either the elemental statements can be viewed as axioms from sets of which various numerical expressions can be derived as theorems, or the numerical expressions like (3a) can be seen as hypotheses about the structure of perception in a given domain and various logically related elemental statements can be treated as predictions from these hypotheses. In this way, specific properties of the dimensional organization of perception are formulated, and testable ordinal consequences of these properties are described which can be compared to the empirically observable order of judged dissimilarity between pairs of vocoids.

Taking any vocoid v_j to be equivalent to its first four formant frequencies ($v_j \equiv [F_{1j}, \dots, F_{4j}]$), several hypotheses about the

dimensional organization and scales for those formant frequencies can be formulated to account for the resulting subjective auditory impression. Expressions (3a) through (3d) below are hypotheses about vocoid perception which range in increasing specificity from (3a) to (3d).

$$(3a) \quad d(v_j, v_k) = \mathbf{G} \left[\phi_1(F_{1j}, F_{1k}), \dots, \phi_4(F_{4j}, F_{4k}) \right] \text{ decomposable}$$

$$(3b) \quad d(v_j, v_k) = \mathbf{G}' \left[|\psi_1(F_{1j}) - \psi_1(F_{1k})|, \dots, |\psi_4(F_{4j}) - \psi_4(F_{4k})| \right] \text{ subtractive}$$

$$(3c) \quad d(v_j, v_k) = \mathbf{G}'' \left[\sum_{i=1}^4 \phi_i(F_{ij}, F_{ik}) \right] \text{ additive}$$

$$(3d) \quad d(v_j, v_k) = \mathbf{G}''' \left[\sum_{i=1}^4 \phi_i |\psi_i(F_{ij}) - \psi_i(F_{ik})| \right] \text{ additive-subtractive}$$

In expressions (3a) - (3d) $\underline{d}$ is a symmetric ordinal distance measure, the $\underline{G}$'s are strictly increasing functions in any of their arguments, ϕ_i are symmetric real-valued functions which equal zero when reflexive ($F_{ij}=F_{ik}$), and ψ_i are psychophysical functions which give interval scale values for F_i . Among expressions (3a) - (3d), (3a) is the most general and least restrictive; it merely requires that each term (here, formant) contribute its effect separately from the others and, if expression (3a) is understood as a hypothesis about vocoid perception, it generates the fewest specific predictions. (3b) and (3c) are specific cases of (3a), and (3d) is the case when (3b) and (3c) both hold. Euclidean distance is one instance of a function of form (3d) and, taken as a hypothesis about vocoid perception, expression (3d) yields a very rich set of ordinal predictions. Our general approach will be to determine how closely the judgments from each listener conform with the patterns of responses predicted by expressions (3a) - (3d) and which properties of the expressions are conformed to or violated by all the listeners. Further, the nature of the disagreements between the listeners' data and the patterns to be expected from various distance functions will be analyzed to isolate the proper algebraic form to satisfactorily express distance between vocoids.

Experiments 1 and 2 examine the testable consequences of expressions (3a) and (3c) to see whether or not all or part of the formant space is perceptually decomposable and interdimensionally additive. Experiment 3 was conducted to identify the causes of some of the disagreements between the listeners' judgments and the predictions from Expressions (3a) - (3c). Experiment 4 was designed to construct interval scales for F1 and F2, and incidentally yielded a rich set of tests for subtractivity in those dimensions. The summary of results (sec. 3.8) discusses alternative dimensional organizations and psychophysical scales that might be suggested by the results of the individual experiments.

3.3 General Experimental Method

3.3.1 Design. Patterns of six formant frequencies were presented. The lowest three formant frequencies were independently varied. The three highest formant frequencies were fixed at 3500, 4500, and 5000 Hz. We sought to determine changes in response patterning associated with manipulation of the first three formants. The experiments elicited sets of comparative judgments necessary for representing vocoid perception in accordance with various algebraic forms. Details of the separate experiments will be given within the 'Experiment' sections. Later experiments typically addressed methodological and substantive issues raised by earlier studies.

3.3.2 Apparatus, Stimuli, and Listeners. Synthetic speech signals were constructed by employing programs and speech synthesizers available to the Speech Communication Group of the Research Laboratory of Electronics at the Massachusetts Institute of Technology. The experimental controller was a PDP-9 computer. The audio stimuli were generated by a terminal analog synthesizer, KLATT; or in one experiment by a parallel formant synthesizer, DICHOT, both developed by Dennis Klatt of MIT and schematically represented in Figure 3.3-1. The synthesizers produced digital files appropriate for digital to analog conversion at 10,000 samples/sec. with 12 bits per sample.

The resulting stimuli were recorded on audio tape with the aid of a program developed by A.W.F. Huggins, GENTAP. Subjects listened to the taped stimuli through SHARPE PRO model HA-660 earphones in a sound proofed listening room. The audio tapes were played on a Presto tape recorder with the level adjusted so that most of the stimuli were between 70 and 80 dB (rel. $.0002 \text{ dynes/cm}^2$) at the listeners' ears.

The invariant properties of the stimuli are shown in Figure 3.3-2. Stimuli were 400 msec. long and all stimuli were completely steady state in their formant frequencies. All stimuli, except in Experiment 1, had fundamental frequency, F_0 , increasing linearly over time from 85 to 115 Hz. The stimuli had a trapezoidal source amplitude envelope as shown in Figure 3.3-2a. Details of stimulus composition are given within the separate experiment sections if the stimuli deviate from these values. It is important to notice that, since the intrinsic relative amplitudes of various F-patterns are quite different, the amplitude and loudness of the stimuli within any one experiment can vary considerably even though the source amplitude is constant within each experiment. For instance, in Experiment 2a, the stimuli vary from 65 to 91 dB (rel. $.0002 \text{ dynes/cm}^2$). Figures 3.3-2b,c specify the constant spectral properties of the 'glottal' pulse train that excited the various F-patterns.

There were six listeners (Ls) engaged in this study, although no more than four were used in any one study. Each L was paid \$3.00/hr and participated one or two hours a day, two or three days a week for a number of weeks. The Ls all reported normal hearing and otherwise they can be described as follows:

<u>L</u>	<u>Age</u>	<u>Sex</u>	<u>Origin (language)</u>
HN	28	M	Mashhad, Iran (heavily accented English)
TB	30	M	Long Island, N.Y.
LL	24	F	Nova Scotia, Canada
RB	28	M	Ankara, Turkey (fluent English)
BI	19	M	Abadan, Iran (fluent English)
JH	25	M	Iowa

3.3.3 Procedure. Signals were recorded in sessions of 100 pairs divided into four sub-blocks of 25 stimulus pairs each. Typically, the first signal was constant over the 100 pairs, and the interval between signals within a pair was 100 msec, with the interval between successive pairs 8 seconds. Successive sub-blocks were separated by 2 minutes. The response format used, in which Ls repeatedly filled out sheets with four columns of 25 spaces, may have facilitated stereotyped responding. However, there were enough different sessions in each experiment that listeners rarely listened to the same session twice during one day. When a particular tape session was repeated in a single day, at least three other sessions intervened. When questioned, some Ls said they were aware of the pattern of the various first pairs in the sub-blocks of some sessions, but other than that they could not remember the pattern of responses that they were making. Also, during Experiments 3 and 4, sessions from both these experiments were usually interspersed for the listener.

Instructions. Listeners were instructed to assign numbers in accordance with the dissimilarity of the two sounds within a pair: large numbers to sounds of large dissimilarity and small numbers to similar sounds. If two sounds were exactly the same, they were to assign zero. They were asked to spread their responses over the range of possible responses, and it was suggested that the range 0-9 would be adequate. Some of the listeners reported that a wider response range would have been preferred, but the highest response recorded was 14 and all listeners except BI responded within 0-9.

A general orientation was given to explain that the sounds were generated by a computer, and although human-like, did not reproduce the sounds of any language. Ls were instructed to try to treat the stimuli as acoustic events completely divorced from any connection with language or orthography. Ls were told to judge the stimuli as they "struck the ear." Each listener then did two or three practice sub-blocks and was asked to repeat back to the experimenter the substance of the instructions and to ask any questions. Some Ls asked

if differences in loudness were important or not, or asked other questions about various qualities of the sounds and their relative importance to the task. The experimenter instructed them that any difference that they could hear (that "struck their ears") was a difference, and that the relative importance of the various kinds of difference in quality was one of the results that was sought. Ls were told that there were no right or wrong answers, but information was wanted about how ears work.

3.3.4 Data Analysis. In most sub-blocks, at least the first three and the last two responses were discarded, in order to avoid stereotyped responding as much as possible within a fixed-order presentation format. Also, for each listener, the responses from the first several runs and the last run on a session were usually discarded. For each pair of vocoids, v_j and v_k , the mean judged magnitude and the standard deviation of the judged magnitudes were obtained for each listener. The number of judgments entering the calculation of the mean and standard deviation differed in the separate experiments and sometimes differed across Ls within an experiment.

We define the confidence interval, C.I. as

$$L_a : CI(v_j, v_k) = \bar{X}_{jk} \pm 2\sigma_{\bar{X}_{jk}}$$

where $\bar{X}_{jk}$ is the mean of L_a 's responses to (v_j, v_k) and $\sigma_{\bar{X}_{jk}}$ is the 'standard error' of the mean. The standard error of the mean is calculated as:

$$\sigma_{\bar{X}} = \frac{\hat{s}}{\sqrt{n}} \quad \text{where} \quad \hat{s} = \sqrt{\frac{\sum (X - \bar{X})^2}{n-1}}$$

here n is the number of responses to the pair from a given listener and $\bar{X}$ is the mean of his magnitude responses.

The measurement theoretic analysis of experimental data only requires ordinal information, so the judged magnitudes for the dissimilarity between pairs of vocoids for a given listener, L_a , are reduced to " $L_a : v_1, v_2 > v_3, v_4$ " [read: for listener L_a the

dissimilarity between v_1 and v_2 is greater than the dissimilarity between v_3 and v_4].

The magnitude ordering of a listener's response to the pairs of vocoids was determined by comparing the calculated CIs. If for two pairs the CIs did not overlap, then the pair of vocoids with the larger mean was accepted as strictly larger.

$$\begin{aligned} &\text{if for } L_a \quad \bar{X}_{12} - 2\sigma_{\bar{X}_{12}} > \bar{X}_{34} + 2\sigma_{\bar{X}_{34}}, \\ &\text{then} \quad L_a : v_1, v_2 > v_3, v_4. \end{aligned}$$

Note that this method of data collection and analysis imposes transitivity on the order relation ">" of the dissimilarities.

Since the definition of ">" admits error and a certain number of false violations could be expected even if listeners operated, for instance, according to Equation (3d) but with some noise, how is one to evaluate the import of a set of violations of a given hypothesis? Violations are likely to indicate a different underlying structure for responses when the violations are systematic and replicable. A set of violations might be systematic in several ways; (1) the pattern might be consistent across Ls or (2) the violations might be delimitable to a small set of stimuli or intervals which are constant in the anomalous directions of their magnitudes or (3) the violations may conform to the expectations from an alternative hypothesis. Violations assume importance in rejecting a hypothesis when they may be constructed at will.

3.4 Experiment 1

Experiment 1 sought to discover which physical attributes of vocoid transfer function spectrum envelopes acted like psychological dimensions in terms of the independence of an attribute's contribution to dissimilarity. Decomposability and additivity (see eq.s (3a) and (3c) above) were taken as conditions for independence. Other purposes of Experiment 1 were (1) to provide general information about the perceptual organization of vocoids in relation to their organization

within formant space, (2) to construct and incorporate a test of the auditory significance of changes in formant intensity levels, (3) to evaluate the relative adequacy of several commonly used psychophysical scales for formant frequency, and (4) to determine how compatible disparate listeners' responses are on this task:

3.4.1 Method.

Listeners. HN and TB.

Stimuli. The F-patterns of the stimuli in Experiment 1 are largely the same as the selection of F-patterns later used in Experiments 2a and 2b, however the invariant properties of the stimuli in Experiment 1 (e.g., the common duration, source amplitude envelope shape, F_0 slope, and spacing in time) were different from these properties in all the other experiments. The values for invariant stimulus properties in Experiment 1 are given in Figure 3.4-1. These values differ from those shown in Figure 3.3-2a.

The 19 F-patterns used in Experiment 1 are listed in Table 3-1 and their locations in formant space are shown in Figure 3.4-2. Table 3-1 lists the equivalent frequencies, in Hertz, for the nominal code employed in Figure 3.4-2. The F-patterns occupy points in a 3x3x2 grid in formant space, with one extra F-pattern at 500, 1500, 2500 (coded as 555 in Fig. 3.4-2 and in Table 3-1). This sample of F-patterns was chosen: (1) to sample the formant space with a uniform logarithmic frequency density; (2) to include both extreme and central vocoids; and (3) to construct critical tests that would be likely to violate (3a) and (3c) (see Fig. 3.4-3 and sec. 3.4.2 "order predictions"). The notation convention for formant frequencies and associated formant bandwidths are given at the top of Figure 3-1.

In Table 3-1, the names of the stimuli correspond to their positions in the three formant representation. "122" can provide an example. In "122" the initial "1" corresponds to the frequency of the first formant: it is at the "1" value, (175 Hz) the lowest frequency among the three first formant frequencies used in these

stimuli. The middle "2" corresponds to the second formant frequency (1200 Hz) which is at the second lowest value out of the three possible second formant frequencies in this stimulus set. The last "2" signifies that the third formant frequency is at the higher of the two possible frequencies (3000 Hz).

Two other notation conventions will be used on occasion for clarity. A subposed dot beneath one of the places in a stimulus name may be used to signify that the indicated formant is constant or forms part of a constant interval in a comparison. For instance, in $21\dot{2} > 12\dot{2}$, the third formant frequency is constant at 1200 Hz. Underlining will be used in Experiment 2a, where the name of a hub stimulus may consist of three numbers, but the numbers have no relation to the nominal code of Table 3-1. For instance, in Section 3.5 we will routinely underline hubs 515 and 510 because in neither of them is F_2 equal to 800, as the middle "1" would suggest.

The representation employed assigns the successive resonant frequencies to successive formants. The lowest, next lowest, and third lowest resonant frequencies become first, second, and third formants respectively. However, because the frequency ranges of F_2 and F_3 overlap, the nominal code is shifted in one circumstance. For example, the name x31 implies: any first formant frequency, nominal code 3 second formant frequency, and nominal code 1 third formant frequency. But, according to Table 3-1, this would indicate an F-pattern (x,2000,1800) which is contrary to the rule that $F_1 < F_2 < F_3$. Thus, x31 is simply (x,1800,2000).

It is noted that in this experiment both stimuli "311" and "312" were mistakenly synthesized as 700, 800, 2300. They were identical as far as visual inspection of a spectrum analysis would show, and they both had a unique F_3 value that invalidates many ordinal comparisons involving them. However, there is barely an audible difference between this sound, 700,800,2300 and either the correct 311 or the correct 312, so that the orderings involving these mis-synthesized stimuli are probably about as indicated. In Experiment 1 both these

stimuli will be referred to as "311.5." In Experiments 2a and 2b, these stimuli are correct.

Procedure. Of the 19 stimuli in Experiment 1, four were chosen to be "hub" stimuli: 555, 112, 221, and 332. Each of these hubs was paired with each of the stimuli including itself. Thus in Experiment 1, there were 76 pairs (4 hubs times 19 stimuli), with the hub stimulus always in the A position in the AB presentations.

One session of 100 pairs was recorded on tape, and in each sub-block of 25 pairs one hub stimulus was paired with each of the 19 stimuli in a quasi-random order. Responses to the first five pairs and to the last pair in a sub-block were discarded, leaving 19 recorded responses out of the 25 pairs in a sub-block. Over the course of Experiment 1, both Ls responded to the taped session 20 times, but all responses from the first five presentations and from the last three presentations of the session were discarded; leaving 12 tabulated responses per listener to each of the 76 pairs.

Data Analysis. For each vocoid pair presented, a sample of 12 judgments per L was recorded. Confidence intervals were calculated and the pairs were ordered as outlined in section 3.3.4.

3.4.2 Design.

To facilitate the analysis of Experiment 1, an interval notation is used. Experiment 1 was designed to test whether or not its three independently controlled variables, F_1 , F_2 , F_3 , behave like dimensions in their contributions to dissimilarity. All of the conditions for dimension-hood outlined by Tversky and Krantz (1970) operate on pairs of frequencies of a formant, and these pairs will be called "intervals." Intervals are the paired input values of the three independent variables; for instance, the interval from 1200 Hz to 1500 Hz along F_2 is symbolized as: $I_2(1200,1500)$. Individual formant frequencies like 1200 and 1500 never enter the design except as endpoints of one or another interval. The dissimilarity of two F-patterns, 175,800,1800

and 350,1200,1800, i.e., $d(111,221)$, can be considered as the combined effect of the two intervals $I_1(175,350)$ and $I_2(800,1200)$ with I_3 constant at (1800,1800). The intervals constructed in the 76 pairs are listed in Table 3-2. For simplicity Table 3-2 leaves out all the reflexive intervals like $I_2(1200,1200)$ and it only lists the 'low-to-high' member of symmetric interval pairs (i.e., $I_2(800,1200)$, but not $I_2(1200,800)$).

Experiment 1 was designed to test the appropriateness of Equation (3d) for expressing the relative auditory dissimilarity among a set of vocoid F-patterns. Equation (3d) requires the independent validity of Equations (3b) and (3c), both of which imply (3a). Furthermore, some functions ψ_i must be specified if an expression of the form (3b) or (3d) is to be applied. As discussed in section 2.2, the validity of some expression like (3d) for vocoids has often been assumed and frequency scales ψ_i have been proposed. Experiment 1 yields no tests of (3b) (and so no test of 3d), but it does test (3a) and (3c), and it contributes toward an empirical solution to the question of formant frequency scaling.

As mentioned in section 3.2, since certain elemental ordinal relations in data are necessary conditions for the valid assertion of general numerical equations like (3a) and (3c), we can treat Equations (3a) and (3c) as hypotheses about the nature of auditory perception and seek to confirm or violate predictions formulated in terms of the order of dissimilarity among pairs of vocoids. Equations (3a) and (3c) imply context-independence in two kinds of elemental predictions; order predictions and equality predictions. Prediction (a), on page 49, is an order prediction. An equality prediction would be of the form: if the only difference (in a formant frequency representation) between a pair of vocoids is a given F2 interval, then the judged dissimilarity of that pair must be equal to the judged dissimilarity of any other pair which differs solely in that same F2 interval. Because the design of Experiment 1 was incomplete and the procedure was insufficiently exact, the tests of equality predictions

are generally not rigorous. Therefore, conclusions are based primarily on the order predictions.

Predictions from decomposability (3a).

Equality Predictions. (3a) predicts that all pairs of F-patterns which differ only in a common interval, or in a common combination of intervals, should be equally dissimilar.

given $IN(r,s) = IN(s,r)$ and $IN(r,r) = IM(p,p)$,

(E1)

then for any three intervals $I1, I2, I3$,

$$I1, I2, I3 = I1, I2, I3$$

For example, consider the vocoid pair $1\dot{1}\dot{2}, 2\dot{1}\dot{2}$ and the pair $2\dot{2}\dot{1}, 1\dot{2}\dot{1}$. Each pair differs only in F1 (nominal code 1 versus code 2) as is indicated by the subposed dots on the constant elements. Under the prediction of single-factor order-independence from decomposability (3a), the dissimilarity between the two members of each pair should be equal. (E1) asserts that composites of equal intervals are equal. By focussing on non-reflexive intervals we may notice the following predictions from (E1): for $I2(800,1200)$ notice $d(1\dot{1}\dot{2}, 1\dot{2}\dot{2}) = d(2\dot{2}\dot{1}, 2\dot{1}\dot{1})$; for $I3(1800,3000)$ notice $d(1\dot{1}\dot{2}, 1\dot{1}\dot{1}) = d(2\dot{2}\dot{1}, 2\dot{2}\dot{2})$; for $I2(800,1200)$ and $I3(1800,3000)$ notice $d(1\dot{1}\dot{2}, 1\dot{2}\dot{1}) = d(2\dot{2}\dot{1}, 2\dot{1}\dot{2})$. These are pairs that must be equal if the auditory quality of vocoids is decomposable into formant frequency components as in Equation (3a).

Order Predictions. (3a) is only valid if all orders of intervals along any one formant frequency are constant regardless of their being combined with any fixed interval or fixed combination of intervals along other formant frequencies (see Tversky and Krantz, 1970, axiom 4).

notation: let $IN, IN'; IM, IM'; IK, IK'$ be six intervals,
two each on FN, FM , and FK .

(E2)

then, $IN, IM, IK \geq IN', IM, IK$ iff $IN, IM', IK' \geq IN', IM', IK'$

Among the orders compared in Experiment 1, some appear to be the most likely candidates to violate (E2). This prediction has been discussed

above and some likely violations of (E2) were incorporated into Experiment 1 to test the decomposability of F2 in relation to F1 and F3. These critical tests were worked out by assuming (Flanagan, 1955, 1957) that formant intensity levels, L_n , or the overall intensity level, L , was perceptually significant, so that a given F2 interval could have very different auditory magnitudes, depending upon the F1 interval with which it is combined. This is contrary to (3a) and hence (E2). The construction of a likely violation of (E2) is illustrated in Figure 3.4-3. Selected examples of critical tests of the form shown in Figure 3.4-3 and found in Experiment 1 are a,b,c,d,e = 555,312, 322,212,222; or 222,311,331,111,131. These tests were designed to detect a violation due to a great increase in level associated with sound "b" (311 or 312). Since 311 and 312 were both mis-synthesized as 311.5, only a subset of the designed tests are realized, and they will require an additional technical assumption (see sec. 3.4.4).

Predictions from additivity (3c).

Any violation of the conditions (E1) and (E2) for (3a) outlined immediately above, is necessarily a violation of (3c). Further, Equation (3c) predicts that orders of joint intervals on any two formant frequencies are constant regardless of any fixed concurrent interval along the remaining formant frequency.

notation: that of (E2)

(E3)

if $(IN, IM, IK) \geq (IN', IM', IK)$
 then $(IN, IM, IK') \geq (IN', IM', IK')$.

Selected examples of (E3) in Experiment 1 are:

I1, I2: $332, 122 \geq 332, 212$ iff $332, 121 \geq 332, 211$

I1, I3: $221, 121 \geq 221, 222$ iff $221, 111 \geq 221, 212$

I2, I3: $555, 121 \geq 555, 132$ iff $555, 221 \geq 555, 232$ iff $555, 321 \geq 555, 332$
 or $112, 121 \geq 112, 132$ iff $112, 221 \geq 112, 232$ iff $112, 321 \geq 112, 332$

There is a great number of apparent tests of additivity in Experiment 1, but because of the mis-synthesis of stimuli 311 and 312 and the alignment of the x31 row, about a third of the possible tests are not

realized. However, Experiment 1 still contains a rich set of tests of the predicted context independence of joint interval orders.

Predictions from frequency scaling.

The last issue that Experiment 1 addresses is the relative adequacy of various commonly used frequency scales to predict the magnitude order of the formant frequency intervals found in the data. Table 3-3a presents four orderings of the single-formant intervals in Experiment 1 that can be placed in ratio-invariant relation to each other without assuming interdimensional additivity. The orderings presented in Table 3-3a are derived from linear, tri-linear, log, and mel frequency scales, and the intervals are displayed at their numerical values.

3.4.3 Results.

a) The mean response magnitude, $\bar{X}$, and the sample standard deviation, s , for each stimulus pair for Listeners HN and TB are given in Table 3-4. The listeners' general agreement at an ordinal level is revealed by inspection of Table 3-4. For instance, the pairs 112,332 and 112,331 were judged most dissimilar and the pair 112,111 was the most similar non-reflexive pair. The average sample standard deviation was smaller for TB (0.73 vs 1.05 for HN). His data, therefore, yielded more ordinal information and, thereby, more inconsistent orderings and violations of predicted order congruencies.

Table 3-5 presents the obtained orders of unidimensional intervals for each listener. For each of the hub stimuli listed at the top of the Table, the consistent ordinal rank-orderings are listed for each listener for each formant. For example, the upper left entry implies that with a hub of 700, 2000, 3000 Hz, the rank ordering of dissimilarity for the first formant for TB was 175 > 350 > 700, all relative to 700.

b) (E2) and (E3) are necessary, but not sufficient conditions, for decomposability and additivity. Further technical conditions must be met to establish decomposability. One such condition, the sym-reflex condition, states that distances must be symmetric and must be minimal

between a stimulus and itself (Tversky and Krantz, 1970, axiom 1).

$$(3.4a) \quad d(j,k) = d(k,j) > d(k,k) = d(j,j)$$

Another condition, weak sequential between-ness, can be stated:

$$(3.4b) \quad \begin{array}{l} \text{for any three values of } F_n; s, t, u, \\ \text{if } s < t < u \\ \text{then } IN(s,t) \leq IN(s,u) \\ \text{and } IN(t,u) \leq IN(s,u). \end{array}$$

Sequential between-ness is needed to construct the allowable tests of (E2) and (E3) involving 311.5 and the x31 stimuli. There are many tests of sym-reflex and sequential between-ness in Experiment 1, the results of which are also covered in Table 3-5. TB gave only one violation of sym-reflex:

$$TB: 221,111 > 221,112 \text{ (i.e., } I3(1800,1800) > I3(1800,3000));$$

and one violation involving either sym-reflex or sequential between-ness:

$$TB: 112,311 > 112,321$$

$$\text{i.e., } I2(800,800), I3(3000,2300) > I2(800,1200), I3(3000,1800).$$

These violations are confounded with the failures of the equality predictions from decomposability. HN gave no violations.

From the last inequality given above, magnitude relations across both frequency and intensity can be constructed; e.g.,

$$IL1(15,57), IL2(14,55) > I2(800,1200), IL1(15,40), IL2(14,35)$$

which can lead to frequency and intensity scales with a common unit. Inequalities of this type will be discussed in section 3.8.

c) Equality predictions from decomposability (3a). Because each sub-block of Experiment 1 was limited to one 'hub' sound in A position in the AB comparisons, the magnitude estimates were not necessarily comparable between hubs. In Table 3-4, several inter-hub distances are asymmetric, which could be due to either stimulus order per se, or to a larger contextual difference. Notice that for both TB and HN $112,332 > 332,112$ and $221,332 > 332,221$, while for TB: $112,221 > 221,112$ and HN: $112,221 = 221,112$. For these reasons and because "="

TABLE 3-1: STIMULI OF EXPERIMENT 1.

Nominal Code for Formant Frequencies and Bandwidths					
	"1"	"2"	"3"		
F1	175	350	700	500	
BW1	80	50	45	50	
F2	800	1200	2000	1500	
BW2	50	60	75	70	
F3	1800	3000		2500	
BW3	65	200		110	

Name	F ₁	F ₂	F ₃	L ₂	L ₁ (rel .0002 dy/cm ²)
Hubs for Experiment 1					
555	500	1500	2500	30	80
112	175	800	3000	18	67
221	350	1200	1800	28	76
332	700	2000	3000	37	83
Other Stimuli for Experiment 1					
111	175	800	1800	19	66
112	175	800	3000	18	67
121	175	1200	1800	16	66
122	175	1200	3000	12	65
131	175	1800	2000	21	68
132	175	2000	3000	12	66
211	350	800	1800	32	79
212	350	800	3000	31	78
221	350	1200	1800	28	76
222	350	1200	3000	24	76
231	350	1800	2000	33	79
232	350	2000	3000	24	77
311	700	800	1800	55	91
312	700	800	3000	54	90
321	700	1200	1800	43	84
322	700	1200	3000	39	82
331	700	1800	2000	46	89
332	700	2000	3000	37	83

** Aurally, and later spectrographically, these sounds were indistinguishable. The spectrograph showed them both to be approximately: 700, 800, 2300, 3500, 4500, in Experiment 1. For later experiments they are correct.

TABLE 3-2: FREQUENCY INTERVALS THAT FORM ELEMENTS IN THE DESIGN OF EXPERIMENT 2a. I1, I2, AND I3 ARE INTERVALS ALONG THE FIRST, SECOND, AND THIRD FORMANT FREQUENCIES RESPECTIVELY. ALL VALUES ARE IN Hz.

I1	I2	I3
(175,350)	(800,1200)	(1800,2000)
(175,500)	(800,1500)	(1800,2300)
(175,700)	(800,1800)	(1800,2500)
(350,500)	(800,2000)	(1800,3000)
(350,700)	(1200,1500)	(2000,2300)
(500,700)	(1200,1800)	(2000,2500)
	(1200,2000)	(2000,3000)
	(1500,1800)	(2300,2500)
	(1500,2000)	(2300,3000)
	(1800,2000)	(2500,3000)

TABLE 3-3a: FREQUENCY INTERVALS AS ORDERED BY FOUR SCALES.

LINEAR 01	TRI-LINEAR 01	LOG 01	MEL 01
175- 175-350	175- 175-350	.9- 800-1200	200- 175-350
277- 350-700	217- 800-1200	1.1- 1200-2000 1800-3000	252- 800-1200
380- 800-1200	260- 350-700	1.3- 175-350 350-700	304- 350-700
482- 175-700	302- 1800-3000 1200-2000	1.5- 800-2000	356- 1200-2000
585- 1200-2000	345- 1800-3000 1200-2000	1.7- 175-700	410- 1800-3000 175-700
687- 1800-3000 800-2000	388- 800-2000	1.9- 175-700	462- 800-2000
790- 1800-3000 800-2000	420- 175-700	2.2- 175-700	514- 800-2000
892- 1800-3000 800-2000	462- 800-2000	2.4- 175-700	566- 800-2000
995- 1800-3000 800-2000	505- 800-2000	2.6- 175-700	620- 800-2000
1097- 1800-3000 800-2000	550- 800-2000	2.8- 175-700	672- 800-2000
1200- 1800-3000 800-2000	600- 800-2000	3.0- 175-700	725- 800-2000

TABLE 3-3b: ORDINAL DATA; TOP TO BOTTOM = SMALL TO LARGE

Subject <u>TB</u>	Subject <u>HN</u>
3000-1800	1200-800, 3000-1800
1200-800	350-175
350-175, 2000-1200	2000-1200
700-350, 2000-800	2000-800, 700-350
700-175	700-175

Combined (As Either/Or)

3000-1800
 1200-800
 350-175
 2000-1200
 700, 350, 2000-800
 700-175

TABLE 3-4: RESULTS FROM EXPERIMENT 1.

Stimuli	n = 12			
	HN		TB	
	Sample Mean $\bar{X}$	Sample S.D. S	Sample Mean $\bar{X}$	Sample S.D. S
555,555	0.08	0.28	0.00	0.00
555,111	5.17	0.69	5.75	0.60
555,112	4.25	1.48	5.75	0.72
555,121	4.58	1.04	4.75	0.72
555,122	3.50	1.12	3.33	0.85
555,132	5.75	1.36	5.33	0.47
555,211	5.58	0.95	5.58	0.86
555,221	4.58	1.26	3.58	0.64
555,212	5.58	0.95	4.33	0.85
555,222	3.58	1.04	2.50	0.87
555,232	5.75	1.16	3.83	0.99
555,321	5.17	1.14	6.00	1.35
555,322	6.25	1.01	6.08	1.04
555,332	8.25	0.72	8.33	0.62
555,131	4.50	0.96	4.25	0.72
555,231	5.33	1.43	3.75	1.01
555,331	7.25	0.83	8.92	0.28
555,311	6.92	1.32	6.92	2.18
555,312	6.17	0.90	7.00	1.00
112,555	5.25	1.01	5.50	0.50
112,111	0.17	0.37	0.00	0.00
112,112	0.00	0.00	0.00	0.00
112,121	0.33	0.47	1.67	0.62
112,122	0.58	0.49	2.08	0.86
112,132	5.08	1.19	5.50	0.50
112,211	3.17	1.14	3.67	0.47
112,221	3.75	0.92	4.25	0.43
112,212	1.92	0.86	3.00	0.41
112,222	1.75	1.01	3.25	0.43
112,232	5.92	1.26	5.42	1.04
112,321	6.33	2.21	7.25	0.43
112,322	6.33	0.94	6.92	0.28
112,332	8.25	0.92	8.83	0.37
112,131	3.92	1.26	4.25	0.83
112,231	4.92	0.86	5.42	0.49
112,331	7.33	0.94	8.92	0.28
112,311	6.67	1.03	8.08	0.28
112,312	6.50	1.04	7.83	0.37

continued

TABLE 3-4: (continued)

n = 12

Stimuli	HN		TB	
	Sample Mean $\bar{X}$	Sample S.D. S	Sample Mean $\bar{X}$	Sample S.D. S
221,555	4.00	1.35	3.75	0.92
221,111	4.00	1.41	5.33	0.74
221,112	3.83	1.34	5.25	0.59
221,121	1.00	0.70	2.33	0.94
221,122	1.25	0.59	4.08	1.18
221,132	4.50	0.95	6.50	0.50
221,211	1.00	0.57	3.75	0.72
221,221	0.08	0.27	0.00	0.00
221,212	1.67	0.62	3.75	0.72
221,222	1.67	0.55	3.08	1.03
221,232	6.17	1.14	5.00	1.15
221,321	5.50	1.04	6.58	0.64
221,322	5.58	0.95	6.92	0.49
221,332	8.00	1.15	8.42	0.64
221,131	4.58	1.11	4.42	0.64
221,231	4.42	1.03	4.42	1.03
221,331	7.00	0.91	8.83	0.37
221,311	5.33	1.17	6.42	2.90
221,312	5.67	0.74	7.33	0.62
332,555	5.33	1.59	5.33	1.24
332,111	6.75	1.08	7.17	0.89
332,112	6.67	1.02	7.25	0.82
332,121	7.00	1.00	7.75	0.82
332,122	5.50	1.80	7.75	0.92
332,132	6.42	1.18	6.17	1.34
332,211	6.25	1.16	5.42	0.64
332,221	6.50	1.11	6.33	1.54
332,212	6.42	0.95	6.33	0.74
332,222	6.08	1.49	6.75	1.01
332,232	6.58	0.86	4.92	0.64
332,321	6.25	1.42	3.33	0.62
332,322	5.00	1.29	2.50	0.76
332,332	0.17	0.37	0.00	0.00
332,131	6.33	1.49	6.83	1.06
332,231	6.17	1.34	5.58	1.11
332,331	3.58	2.05	1.58	0.49
332,311	7.00	1.22	3.33	0.47
332,312	6.92	1.11	3.42	0.49

TABLE 3-5: ORDERS OF UNIDIMENSIONAL INTERVALS IN EXPERIMENT 1.

HUB →		332	221	112	555
		700,2000,3000	350,1200,1800	175,800,3000	500,1500,2500
FORMANT ↓ F ₁	TB	175>350>700	700>175>350	700>350>175	700>175>350
	HN	175>700 350>700	700>175>350	700>350>175	700>175 700>350
F ₂	TB	800>1200>2000	2000>800>1200	2000>1200>800 2000>800>1200	2000>800>1200 800>2000>1200
	HN	800>1200>2000	2000>800>1200	2000>1200>800	2000>800>1200
F ₃	TB	1800>3000	3000>1800	1800>3000	3000 1800 1800 3000
	HN	1800>3000	3000>1800	1800>3000	3000>1800 1800>3000

Order of differences between hubs and other stimuli, by formant and listener, with single formant frequency differences and constant F-pattern otherwise.

For instance: The numbers in the top right-hand box indicate that for TB $I1(500,700) > I1(500,175) > I1(500,350)$ and for HN $I1(500,700) > I1(500,175)$ and $I1(500,700) > I1(500,350)$. The lower left-hand box indicates that for both TB and HN $I3(3000,1800) > I3(3000,3000)$.

has not been defined here, there are no rigorous tests of equality predictions (E1) from decomposability, but there are eight indicative tests of the type schematized in Figure 3.4-4.

HN violated the equality predictions in only one of the eight tests - HN: $221,122 < 112,211$ - which indicates a different magnitude for $I_1(350,175), I_3(1800,3000)$ as opposed to $I_1(175,350), I_3(3000,1800)$ in the context of $I_2(1200,1200)$ versus $I_2(800,800)$. TB indicated violations in four of the eight tests. One violation was TB: $221,111 > 112,222$. This violation is like HN's only equality violation in that both are consistent with a hypothesis that a given F_1 interval is greater when the L_1 's are greater - as they are (slightly) when F_2 or F_3 is lower, other factors constant. The remaining indicative violations from TB are:

(given $221,112 = 112,221$)* *=" undefined
 $221,211 > 112,122$
 $221,222 > 112,111$
 $221,212 > 112,121$.

All three of which are consistent with a hypothesis that formant frequency intervals sound larger when the formant levels are greater as can be seen by inspecting the last two columns of the lower half of Table 3-1. L_2 is 10-15 dB greater when $F_1 = 350$ than when $F_1 = 175$. TB's violations all indicate that fixed (symmetric) F_2 and F_3 intervals are judged greater when heard in a context of higher F_1 .

d) Order predictions from decomposability (3a). If, as indicated above, the equality predictions from decomposability are invalid, then violations of the elemental order predictions (E2) from decomposability are equivocal. However, among a large number of rigorous tests available in the design of Experiment 1, neither HN nor TB produced any violations of (E2) except TB's order reversals in Table 3-5 which are discussed above in section 3.4.3b.

e) Order predictions from additivity (3c). The order predictions (E3) from additivity may also be equivocal due to the indicated failure

of unidimensional intervals to have constant magnitude. More than half of the available tests of (E3) in Experiment 1 are for distances out from hub 555. Because sequential between-ness had to be assumed to construct some of the tests involving 311.5 and x31 sounds, the number of tests will be counted directionally. In this manner;

"if $555,332 < 555,222$ then $555,331 < 555,221$ "

counts as one test, whereas

" $555,322 \leq 555,112$ iff $555,321 \leq 555,111$ "

counts as two tests. Although either kind of test permits only one violation, the 'two-way' test is a priori twice as likely to produce a violation.

There were 27 tests of joint-order independence (E3) for I1,I2 in Experiment 1, and neither listener gave any inconsistent orders, i.e., neither listener violated the predictions from additivity for F1,F2. There were 28 tests of (E3) for I1,I3, and neither listener gave any violations in these tests either. Among 90 tests of (E3) for joint intervals I2,I3, TB's data yielded only one quadruple of pairs that violated additivity; TB: $555,131 < 555,132 :: 555,331 > 555,332$ and these orders are not very robust. HN did not give a violation on this quadruple, and there were no violations from either listener among the other 89 tests of I2,I3 order independence.

f) With (E2) and (E3) generally satisfied in Experiment 1, it was fair to assume that frequency scales could be constructed for F1, F2, and F3. The relative magnitude of intervals are given in Table 3-3b separately for each listener and then combined. The individual orders or the combined order can be compared to the orders derived from the four common frequency scales displayed in Table 3-3a, and it seems by any obvious measure that a logarithmic scale yields the best fit to the data, tri-linear is nearly as good, and mel and linear scales are far off. No commonly used scale correctly predicts all orders of unidimensional intervals.

g) Summary. The listeners' data were in close agreement and conformed to most of the structures predicted by decomposability and

additivity, while failing to satisfy technical conditions. Results indicated a tendency for a given frequency interval, IN , to be augmented by an increase in intensity level, L_n ; which would violate decomposability. The logarithmic frequency scale best predicted the magnitude order of the frequency intervals.

3.5 Experiment 2

Experiment 2 was run in two parts, 2a and 2b. Experiment 2a was very similar in design to Experiment 1, while incorporating several improvements. The method, design, and results of Experiment 2a are discussed below in sections 3.5.1 through 3.5.3. Experiment 2b was a small extension of Experiment 2a that specifically constructed violations of (F_1, F_2, F_3) decomposability unavailable in Experiments 1 and 2a. The design, method, and results of Experiment 2b are presented in sections 3.5.4 through 3.5.6 and a results summary for 2a and 2b is given in section 3.5.7.

3.5.1 Method

Listeners. LL, BI, RB (see sec. 3.3.2).

Stimuli. Figure 3.3-2 shows the constant source and temporal properties of the stimuli in Experiment 2a. The stimuli were constructed as described in section 3.3.2. Of 25 stimuli used in Experiment 2a, 18 had the F-patterns for the 'other stimuli' 111, 112, 121, ..., 331, 332 listed in Table 3-1. The stimuli will be referred to according to the nominal code at the top of Table 3-1. The nominal code is explained in section 3.4.1, except that in Experiment 2a a fourth value of F_2 is used. Nominal code "4" on F_2 is equivalent to 2500 Hz. Note also, that the names of the hub stimuli used in Experiment 2a are underlined, indicating that they do not follow the nominal code of Table 3-1. Again, a subposed dot may be used to indicate that a certain formant frequency is constant or forms part of a constant interval.

The sounds 311 and 312 were correctly re-synthesized for Experiment 2a. The four 'hub' stimuli in Experiment 2a were:

Name	F-pattern	L ₂
<u>515</u>	<u>500</u> , <u>1500</u> , 2500	30
<u>510</u>	<u>500</u> , <u>1000</u> , 2500	33
<u>2515</u>	<u>250</u> , <u>1500</u> , 2500	17
<u>2510</u>	<u>250</u> , <u>1000</u> , 2500	19

Three additional stimuli included in Experiment 2a were:

Name	F-pattern	L ₂
142	175, 2500, 3000	19
242	350, 2500, 3000	31
342	700, 2500, 3000	43

Procedure. The apparatus is described in section 3.3.2. The instructions and response format are described in section 3.3.3. The four hubs of Experiment 2a were each combined with all the 21 other stimuli forming 86 different pairs presented to listeners. In order to lessen the influence of order effects on the data, each group of 21 pairs between a hub and the other 21 stimuli was presented to listeners in four different quasi-random orders, thus comprising four 100-pair sessions in Experiment 2a. Each session had one hub constantly in the A position of the 100 AB pairs and each session included four sub-blocks of 25 pairs. Each sub-block included a different quasi-random order of a given hub with the other 21 stimuli. Thus, each session yielded four judgments for each of 21 pairs including a particular hub. Each listener responded to each tape session at least five times (LL responded to the tapes eight times each), yielding, after practice, 12 tabulated responses per pair from listeners BI and RB, and 24 tabulated responses per pair from LL.

Data Analysis. Confidence intervals were calculated about the mean judged magnitude for each pair from each listener, and the pairs were ordered for the listeners as outlined in section 3.3.4.

3.5.2 Design

Like Experiment 1, Experiment 2a yielded tests of Equations (3a) and (3c), interdimensional decomposability and additivity. Section

3.4.2 describes two elemental ordinal relations, (E2) and (E3), that should inhere in the data from Experiment 2a (like Experiment 1) if vocoid psychoacoustics can be modeled as a decomposable and additive combination of F_1, F_2, F_3 .

The relation of the hubs to the other stimuli in Experiment 2a is pictured in Figure 3.5-1. Since all four hubs in Experiment 2a are out of line with the 'grid' formed by the other stimuli (see Figure 3.5-1), Experiment 2a included a very rich set of tests of inter-dimensional additivity, whereas Experiment 1 had few such tests. Among the 84 distances judged by the listeners, there are about 2000 instances of distance quadruples that fit (E3) in each of the three dimension pairs. That is, there are about 2000 joint interval order independence tests in F_1, F_2 versus F_3 , and there is a like number of joint interval tests on F_1, F_3 versus F_2 and on F_2, F_3 versus F_1 . The following is just one example of a recondite F_2, F_3 additivity test that is included in Experiment 2a:

$$\underline{515,222} > \underline{510,211} \quad \text{iff} \quad \underline{2515,322} > \underline{2510,311}$$

However, there are actually considerably fewer than 6000 tests of additivity in Experiment 2a because any test of decomposability, for instance in F_2

$$\underline{515,211} > \underline{515,221} \quad \text{iff} \quad \underline{2515,111} > \underline{2515,121}$$

counts as a test of additivity in the joint order of both F_1, F_2 and F_2, F_3 .

3.5.3 Results

Table 3-6 presents the mean response magnitude, $\bar{X}$, and the sample standard deviation, s , for each of the 84 stimulus pairs for each of the three listeners. The general ordinal agreement among the listeners' responses is evident.

Table 3-7a,b presents the results of Experiment 2a in relation to the ordinal predictions (E3) of interdimensional additivity (3c). According to the schema shown in Table 3-7a the tests of (E3) are

TABLE 3-6: RESULTS FROM EXPERIMENT 2a.

Listener	BI (N=12)		LL (N=24)		RB (N=12)	
	$\bar{X}$	S	$\bar{X}$	S	$\bar{X}$	S
Stimulus						
515,111	4.50	0.76	3.54	0.57	3.17	0.68
515,112	4.58	0.64	3.25	0.82	3.33	0.74
515,121	4.83	0.37	2.92	0.86	3.08	0.75
515,122	4.75	0.43	3.17	0.89	3.42	0.86
515,131	4.00	0.70	3.79	1.04	4.17	0.79
515,132	5.67	0.47	4.87	0.43	4.17	0.68
515,142	5.83	0.37	5.54	0.57	4.50	0.86
515,211	4.67	0.62	4.83	0.79	3.92	0.64
515,212	4.58	0.49	4.63	0.94	4.08	0.75
515,221	4.42	0.64	3.21	0.70	2.50	0.64
515,222	3.67	0.74	3.17	1.06	4.08	0.75
515,231	4.33	1.10	4.92	1.44	4.00	1.77
515,232	5.08	0.27	5.00	1.22	5.08	0.95
515,242	6.33	0.62	6.13	0.78	6.42	0.64
515,311	5.83	0.37	7.58	0.70	7.17	0.55
515,312	5.92	0.27	7.54	0.49	6.83	0.55
515,321	5.83	0.37	6.83	0.84	6.17	0.55
515,322	5.67	0.47	5.54	0.95	5.83	0.55
515,331	6.42	0.95	7.42	0.70	7.92	0.64
515,332	6.92	1.11	7.38	0.69	7.58	0.49
515,342	9.58	0.75	8.75	0.43	8.67	0.47
510,111	4.83	0.37	3.29	0.61	3.08	0.27
510,112	4.83	0.37	3.29	0.67	3.00	0.40
510,121	4.75	0.43	3.21	0.64	3.00	0.40
510,122	4.83	0.37	3.21	0.70	2.75	0.59
510,131	4.67	0.62	4.58	0.95	4.33	0.62
510,132	5.75	0.43	5.00	0.50	4.17	0.68
510,142	5.92	0.27	5.63	0.48	4.75	0.92
510,211	4.17	0.89	3.42	1.44	2.50	0.86
510,212	3.75	0.72	3.46	1.25	2.33	0.62
510,221	3.67	0.62	3.13	0.83	4.33	0.62
510,222	3.67	0.74	3.17	1.10	3.83	0.68
510,231	4.92	0.64	5.33	1.37	4.92	1.04
510,232	5.25	0.43	4.92	1.11	5.33	0.75
510,242	6.08	0.64	6.08	1.28	6.58	0.49
510,311	6.08	0.27	7.83	0.37	7.00	0.70
510,312	5.92	0.27	7.46	0.70	7.00	0.81
510,321	5.92	0.27	6.92	0.64	5.92	0.64
510,322	5.83	0.37	5.46	1.07	5.92	0.49
510,331	6.92	0.64	7.67	0.62	7.83	0.68
510,332	7.33	0.74	7.25	0.72	7.75	0.43
510,342	9.67	0.62	8.63	0.56	8.66	0.47

TABLE 3-6: (continued)

Listener	BI (N=12)		LL (N=24)		RB (N=12)	
	$\bar{X}$	S	$\bar{X}$	S	$\bar{X}$	S
Stimuli						
2515,111	3.17	0.89	2.29	0.45	3.42	0.75
2515,112	3.25	0.82	1.96	0.53	2.92	0.49
2515,121	3.50	0.64	2.29	0.53	2.67	0.47
2515,122	3.17	0.55	1.92	0.49	2.67	0.47
2515,131	4.75	0.92	3.92	0.81	2.92	0.86
2515,132	5.33	0.62	4.67	0.74	3.50	0.64
2515,142	5.58	0.49	5.55	0.57	4.17	0.68
2515,211	5.08	0.27	4.50	0.76	4.33	0.47
2515,212	5.08	0.27	4.42	0.75	4.00	0.57
2515,221	4.75	0.72	3.25	0.59	4.00	0.70
2515,222	4.58	0.49	3.17	0.37	3.75	0.59
2515,231	5.92	0.49	5.70	0.88	5.25	0.72
2515,232	5.75	0.59	5.20	0.64	5.75	0.92
2515,242	6.75	0.59	6.29	0.67	6.42	0.64
2515,311	5.92	0.49	8.04	0.61	7.00	0.00
2515,312	5.83	0.37	7.70	0.73	7.08	0.49
2515,321	5.92	0.27	6.91	0.81	5.83	0.37
2515,322	5.75	0.43	5.29	0.73	5.83	0.68
2515,331	7.25	0.72	7.75	0.59	8.00	0.91
2515,332	8.08	0.64	7.29	0.61	7.75	0.59
2515,342	10.08	0.27	8.67	0.62	8.83	0.37
2510,111	2.67	1.10	2.20	0.49	2.75	0.43
2510,112	2.50	1.32	2.08	0.86	2.58	0.49
2510,121	2.83	0.98	2.20	0.76	2.67	0.47
2510,122	2.33	1.31	2.00	0.57	2.33	0.47
2510,131	4.75	0.72	4.17	0.79	4.00	0.70
2510,132	5.08	0.49	4.83	0.74	3.75	0.59
2510,142	5.50	0.50	5.58	0.57	4.42	0.75
2510,211	5.00	0.00	4.25	0.52	4.33	0.62
2510,212	5.00	0.40	4.17	0.74	4.00	0.70
2510,221	5.33	0.74	3.42	0.64	4.50	0.50
2510,222	5.00	0.81	3.38	0.69	3.75	0.92
2510,231	5.75	0.72	5.67	0.98	5.50	0.86
2510,232	5.75	0.72	5.50	0.70	5.50	0.76
2510,242	6.50	0.50	6.29	1.05	6.42	0.64
2510,311	6.33	0.62	8.00	0.40	7.08	0.49
2510,312	5.92	0.27	7.58	0.70	6.91	0.75
2510,321	5.75	0.59	7.00	0.57	6.17	0.55
2510,322	6.00	0.40	5.70	0.78	5.83	0.68
2510,331	7.75	0.43	7.41	0.75	8.08	0.64
2510,332	7.83	1.28	7.17	0.74	7.83	0.55
2510,342	10.58	1.03	8.54	0.49	8.83	0.37

TABLE 3-7a: DEFINITIONS.

	_____	_____ :: _____	_____
Confirmation	>		>
Violation	>		<
No Test	>		~
No Test	~		~

TABLE 3-7b: RESULTS OF EXPERIMENT 2a.

	% Confirmation / % No-Test / % Violation		
	LL	RB	BI
F_1, F_2 (260 tests)	85/15/0	85/15/0	45/55/0
F_1, F_3 (1700 tests)	25/75/0	45/55/1 (8)	30/70/0
F_2, F_3 (1700 tests)	55/45/1 (13)	60/40/1 (11)	60/40/1 (12)

% C / % NT / % V

TABLE 3-8: DISTANCE QUADRUPLES THAT VIOLATE DECOMPOSABILITY AND ADDITIVITY IN EXPERIMENT 2a.

<u>RB's</u>	Violations of Decomposability in F_2		
	$\underline{555},222 > \underline{510},212$	::	$\underline{515},322 < \underline{510},312$
	$\underline{515},222 > \underline{510},212$	::	$\underline{2515},322 < \underline{2510},312$
	$\underline{510},221 > \underline{510},211$	::	$\underline{510},321 < \underline{510},311$
	$\underline{510},221 > \underline{510},211$	::	$\underline{2510},321 < \underline{2510},311$
	$\underline{510},222 > \underline{510},212$	::	$\underline{510},322 < \underline{510},312$
	$\underline{510},222 > \underline{510},212$	::	$\underline{2510},322 < \underline{2510},312$
<u>RB's</u>	Violations of Decomposability in F_1		
	$\underline{515},242 > \underline{515},142$	::	$\underline{510},212 < \underline{510},122$
	$\underline{510},112 > \underline{510},212$	::	$\underline{510},122 < \underline{510},222$
	$\underline{510},112 > \underline{510},212$	::	$\underline{510},132 < \underline{510},232$
	$\underline{510},112 > \underline{510},212$	::	$\underline{510},142 < \underline{510},242$
<u>RB's</u>	Violations of Additivity in F_1, F_3		
	$\underline{515},212 > \underline{515},111$	::	$\underline{510},212 < \underline{510},111$
	$\underline{515},122 > \underline{515},221$	::	$\underline{510},122 < \underline{510},221$
	$\underline{515},222 > \underline{515},121$	::	$\underline{510},212 < \underline{510},111$
	$\underline{510},111 > \underline{510},212$	::	$\underline{510},121 < \underline{510},222$
<u>RB's</u>	Violations of Additivity in F_2, F_3		
	$\underline{515},222 > \underline{510},211$	::	$\underline{515},322 < \underline{510},311$
	$\underline{515},222 > \underline{510},211$	::	$\underline{2515},322 < \underline{2510},311$
	$\underline{510},221 > \underline{510},212$	::	$\underline{510},321 < \underline{510},312$
	$\underline{510},222 > \underline{510},211$	::	$\underline{510},322 < \underline{510},311$
	$\underline{510},222 > \underline{510},211$	::	$\underline{2510},322 < \underline{2510},311$

TABLE 3-8: (continued)

BI's Violations of Decomposability in F_2

<u>515</u> ,121 > <u>515</u> ,131	::	<u>2515</u> ,121 < <u>2515</u> ,131
<u>515</u> ,121 > <u>515</u> ,131	::	<u>2515</u> ,221 < <u>2515</u> ,231
<u>515</u> ,121 > <u>515</u> ,131	::	<u>2515</u> ,321 < <u>2515</u> ,331
<u>510</u> ,111 > <u>515</u> ,131	::	<u>2510</u> ,111 < <u>2515</u> ,131
<u>510</u> ,111 > <u>515</u> ,131	::	<u>2510</u> ,211 < <u>2515</u> ,231
<u>510</u> ,111 > <u>515</u> ,131	::	<u>2510</u> ,311 < <u>2515</u> ,331

BI's Violations of Additivity in F_2, F_3

<u>510</u> ,112 > <u>515</u> ,131	::	<u>2510</u> ,112 < <u>2515</u> ,131
<u>510</u> ,112 > <u>515</u> ,131	::	<u>2510</u> ,212 < <u>2515</u> ,231
<u>510</u> ,112 > <u>515</u> ,131	::	<u>2510</u> ,312 < <u>2515</u> ,331
<u>510</u> ,122 > <u>515</u> ,131	::	<u>2510</u> ,122 < <u>2515</u> ,131
<u>510</u> ,122 > <u>515</u> ,131	::	<u>2510</u> ,222 < <u>2515</u> ,231
<u>510</u> ,122 > <u>515</u> ,131	::	<u>2510</u> ,322 < <u>2515</u> ,331

LL's Violations of Decomposability in F_2

<u>510</u> ,231 > <u>510</u> ,211	::	<u>510</u> ,331 < <u>510</u> ,311
<u>2510</u> ,231 > <u>2510</u> ,211	::	<u>2510</u> ,331 < <u>2510</u> ,311

LL's Violations of Additivity in F_2, F_3

<u>515</u> ,132 > <u>515</u> ,111	::	<u>2515</u> ,332 < <u>2515</u> ,311
<u>515</u> ,132 > <u>510</u> ,111	::	<u>2515</u> ,332 < <u>2510</u> ,311
<u>510</u> ,132 > <u>515</u> ,111	::	<u>2510</u> ,332 < <u>2515</u> ,311
<u>510</u> ,132 > <u>510</u> ,111	::	<u>2510</u> ,332 < <u>2510</u> ,311
<u>2515</u> ,132 > <u>2515</u> ,111	::	<u>2515</u> ,332 < <u>2515</u> ,311
<u>2515</u> ,132 > <u>2510</u> ,111	::	<u>2515</u> ,332 < <u>2510</u> ,311
<u>2515</u> ,232 > <u>2510</u> ,211	::	<u>2515</u> ,332 < <u>2510</u> ,311
<u>2510</u> ,132 > <u>2515</u> ,111	::	<u>2510</u> ,332 < <u>2515</u> ,311
<u>2510</u> ,232 > <u>2515</u> ,211	::	<u>2510</u> ,332 < <u>2515</u> ,311
<u>2510</u> ,132 > <u>2510</u> ,111	::	<u>2510</u> ,332 < <u>2510</u> ,311
<u>2510</u> ,232 > <u>2510</u> ,211	::	<u>2510</u> ,332 < <u>2510</u> ,311

categorized as 'confirmations', 'violations', and 'no tests.' Among the tests involving orders of F_1, F_2 joint intervals only those that compose four distances from a common hub were checked. Thus, only about 15% (260) of the possible instances of (E3) on F_1, F_2 were tested. All of the tests for joint interval order independence (E3) in F_1, F_3 and F_2, F_3 were checked, and the percentages of confirmations, violations, and no-tests for each subject for each group of tests is shown in Table 3-7b.

The figures in Table 3-7b may be read, for instance, from the left-middle box "25/75/0": listener LL's data confirmed F_1, F_3 interdimensional additivity (3c) in one quarter of the 1700 tests, her data were too noisy to form strong tests in three quarters of the cases checked and no set of distances that formed a test of (E3) was ordered in violation of this prediction from (3c). The numbers in parentheses below the percentages are the absolute number of distance quadruples in the data that violated (E3). The most important feature of these violations (which cannot be read from Table 3-7) is the inconsistency across listeners (see Table 3-8). Although all three listeners produced about a dozen violations of F_2, F_3 additivity, they form a total of 36 different violations - no listener showed the same violation as any other.

Table 3-7 then can be interpreted to say that in about half the tests of interdimensional additivity, listeners' data were clean enough to form strong tests and among those tests, fewer than one percent formed violations of the orderings predicted by Equation (3c). However, closer inspection of the data shows that although the stimuli used might have formed many tests to check the perceptual salience of L_2 differences in the context of F_1, F_2, F_3 differences as outlined in section 3.4.2 and Figure 3.4-3, the stimuli actually formed very few such tests. The employed stimuli could not generally produce F_1, F_2, F_3 incongruence resulting from the 'perceptual intrusion' of L_2 , particularly in relation to the stimuli 311 and 312 being judged as 'very different.' Figure 3.5-2 displays the stimuli of Experiment 2a

approximately as they would be represented in an F_1 by F_2 plot roughly scaled to fit the response magnitude to distances from hubs 515 and 2510 to the other stimuli to which 311 and 312 were to be compared. This figure indicates why the choice of stimuli in Experiment 2a offered so few violations of decomposability even though distances to 311 and 312 were much greater than would be expected from the proportions of Figure 3.5-2. Since $I_1(500,700)$ was always much larger than $I_1(350,500)$ and $I_2(1000,1200)$ was rarely smaller than $I_2(800,1000)$, the great distance of 311 and 312 from hubs 515, 510, and 2510 would not generally result in incongruent, violating orders. The data from RB, however, did include a small but reliable difference, $I_2(1000,1200) > I_2(800,1200)$, so that the disproportionately large distances to 311 and 312 did violate (E2) in his data (see Table 3-8).

The violations of joint I_2, I_3 order independence from listeners LL and RB all involve 311 and 312 being judged as relatively too far from the hubs. But again, the violations form no compelling pattern across listeners.

From these results, especially as shown in Table 3-7, it would seem that the effects of formant level on vocoid perception may be too small to show up reliably in an experimental design that chooses stimuli on an a priori basis. Perhaps when the formant levels are non-monotonically related to formant frequencies the changes in loudness are too small to produce salient ordinal incongruence in the context of great F-pattern differences. Certainly, the mere 44 violations of interdimensional additivity in F_1, F_2, F_3 as found in Experiment 2a (i.e., less than 1% of the tests) are neither numerous enough nor systematic enough to reject a hypothesized inter-vocoid auditory distance of a form like Equation (3c).

3.5.4 Experiment 2b (design)

Experiment 2b attempted to demonstrate that violations of decomposability in an F_1, F_2, F_3 representation of vocoids could be constructed at will by associating given F_n intervals with different dependent

L_n intervals through changes in F_m ($m \neq n$). Experiment 2a showed that a definite magnitude difference between instances of an interval, e.g., $I_2(800,1000)$, need not produce any ordinal level violations. If $I_1(x,700)$ and $I_2(800,y)$ are consistently the largest perceptual intervals in F_1 and F_2 , then no matter how disproportionately distant $(700,800,z)$ seems from hub (x,y) , no ordinal violation will be realized.

Experiment 2b introduced small F_1 and F_2 intervals which included 700 Hz and 800 Hz as end points. This was accomplished by redoing a part of Experiment 2a, but with the hub stimuli re-aligned so that hub F_1 was shifted up toward 700 Hz and hub F_2 was shifted down toward 800 Hz.

Experiment 2b also included tests on another small group of new stimuli in the $F_1 = 400$ plane. These stimuli lay near the $F_2 = F_3$ line, and were chosen in an attempt to construct violations of decomposability dependent on changes in L_2 and/or in L_3 .

Design. Experiment 2b had two parts: one testing single factor independence (decomposability) for F_1 and F_2 ; and the other testing single factor independence for F_2 and F_3 . Experiment 2b, as analyzed, only included tests of (E2). The design reflected a more careful attempt to construct the kind of violation shown in Figure 3.4-3 for stimuli near $F_1 = F_2$ or near $F_2 = F_3$.

In interval terms, the first part of Experiment 2b ordered $I_1(350,625)$ and $I_1(625,700)$ as combined with six different concurrent I_2 's. In the F_2 factor, the six I_2 's (800,925), (800,1600), (925,1200), (925,2000), (1200,1600), and (1600,2000) were ordered, while combined with $I_1(300,350)$, $I_1(300,700)$, $I_1(350,625)$, or $I_1(625,700)$. All I_1, I_2 comparisons were accompanied by an $I_3(2300,3000)$, since the hubs were in $F_3 = 2300$ plane and all the other stimuli were in $F_3 = 3000$. As in Figure 3.4-3, violations were expected from comparisons which would show (700,800,3000) to be "too far away."

The second part of Experiment 2b paired intervals in F_2 and F_3 . $I_2(1000,1150)$ and $I_2(1150,1300)$ were crossed with $I_3(1500,1700)$ and $I_3(1700,2100)$. Also $I_2(1300,1600)$ and $I_2(1600,1900)$ were crossed

with I3(2100,2500), and I3(2500,2800). In part 2 of Experiment 2b, all stimuli had $F_1 = 400$ Hz, so I1(400,400) was constant, and violations were expected from disproportionately large judgments involving (400,1300,1500) and (400,1900,2100).

3.5.5 Method

Listeners. LL, BI, RB.

Stimuli. In part one, the stimuli were 212,222,232,312,322, and 332 from Experiment 2a, paired with the following hubs with bandwidths $B_1 = 50$, $B_2 = 70$, and $B_3 = 110$.

Name	F-pattern	L_2
26	625,1600,2300	36
27	625,925,2300	41
28	300,925,2300	24

Table 3-9 lists the acoustic characteristics of the stimuli in part two.

Procedure. Three taped sessions were used in the first part of Experiment 2b. One quasi-random order of 100 B position stimuli divided into two sub-blocks of 50 was paired with hubs 26 and 27 and resulted in two sessions. Each sub-block of 50 included four scored instances of each of 12 stimuli; the six xx2's listed above and the six corresponding xx1's. The first and twenty-sixth pairs were included as unscored extras to head the columns in the score sheet. The third session in part one was an 80 pair session divided into two sub-blocks of 40. These sub-blocks were two orders of the eight stimuli, 212, 222,312, and 322, and their xx1 counterparts paired with hub 28 scored four times per sub-block.

In the second part of Experiment 2b, there was one session, divided into four sub-blocks of 25 pairs. Each sub-block paired one hub or the other with each of the eight stimuli in Table 3-9. Each sub-block started with a pair 8,1 or 1,8 (as appropriate) and included

TABLE 3-9: STIMULI USED IN THE SECOND PART OF EXPERIMENT 2b.

All Stimuli: $F_1 = 400$, $B_1 = 50$
 $B_2 = 70$
 $B_3 = 110$

	Name	F_2	F_3	L_2	L_3
(hub)	1	1150	1700	30	16
	2	1000	2100	29	7
	3	1900	2100	36	31
	4	1300	2100	27	14
	5	1300	1500	35	27
	6	1300	2800	25	14
	7	1900	2800	27	24
(hub)	8	1600	2500	26	18
	9	1000	1500	32	16

TABLE 3-10: DATA FROM EXPERIMENT 2b. (Stimulus names explained in Tables 3-1, 3-9, and section 3.5.5.)

Listener	LL		BI		RB	
Practice n	(6)		(0)		(4)	
N	18		16		12	
Stimuli	$\bar{X}$	$2\sigma_{\bar{X}}$	$\bar{X}$	$2\sigma_{\bar{X}}$	$\bar{X}$	$2\sigma_{\bar{X}}$
26,212	5.33	0.39	4.94	0.12	5.33	0.37
26,222	3.72	0.50	5.06	0.22	5.33	0.37
26,232	5.00	0.39	5.25	0.22	4.41	0.45
26,312	7.44	0.43	6.31	0.23	6.08	0.29
26,322	4.39	0.40	6.25	0.22	4.75	0.50
26,332	7.11	0.22	6.63	0.92	5.33	0.28
Practice n	(6)		(0)		(4)	
N	18		16		12	
Stimuli	$\bar{X}$	$2\sigma_{\bar{X}}$	$\bar{X}$	$2\sigma_{\bar{X}}$	$\bar{X}$	$2\sigma_{\bar{X}}$
27,212	5.00	0.36	5.00	0.00	4.92	0.16
27,222	3.61	0.36	5.56	0.25	5.08	0.16
27,232	5.78	0.20	5.81	0.27	5.25	0.26
27,312	2.06	0.11	2.25	0.69	2.91	0.29
27,322	4.33	0.45	2.69	0.53	3.00	0.60
27,332	7.11	0.15	8.19	0.20	6.33	0.28
Practice n	(12)		(0)		(0)	
N	18		10		10	
Stimuli	$\bar{X}$	$2\sigma_{\bar{X}}$	$\bar{X}$	$2\sigma_{\bar{X}}$	$\bar{X}$	$2\sigma_{\bar{X}}$
28,212	2.67	0.36	4.90	0.20	2.80	0.26
28,222	3.61	0.28	4.40	0.61	3.60	0.44
28,312	7.56	0.29	7.00	0.29	6.20	0.49
28,322	5.17	0.18	6.90	0.62	5.40	0.32
Practice n	(12)		(0)		(2)	
N	18		12		10	
Stimuli	$\bar{X}$	$2\sigma_{\bar{X}}$	$\bar{X}$	$2\sigma_{\bar{X}}$	$\bar{X}$	$2\sigma_{\bar{X}}$
1,2	3.33	0.51	4.00	0.55	3.80	0.58
1,4	2.17	0.43	3.33	0.51	3.20	0.40
1,5	2.61	0.62	3.00	0.69	2.90	0.55
1,9	1.50	0.29	2.92	0.75	3.20	0.40
Practice n	(12)		(0)		(2)	
N	18		12		10	
Stimuli	$\bar{X}$	$2\sigma_{\bar{X}}$	$\bar{X}$	$2\sigma_{\bar{X}}$	$\bar{X}$	$2\sigma_{\bar{X}}$
8,3	6.83	0.33	4.83	1.15	6.20	0.26
8,4	4.00	0.16	5.08	0.16	3.40	0.32
8,6	3.72	0.53	4.83	0.59	3.60	0.53
8,7	3.50	0.63	3.42	0.57	3.90	0.86

three scored instances of eight pairs in a different order. The first two sub-blocks had hub 1 in A position; the second two had hub 8 in A position.

Data Analysis. Confidence intervals were calculated and orders, ">", of dissimilarity derived as in section 3.3.4.

3.5.6 Results

The means, $\bar{X}$, and twice the standard error, $2\sigma_{\bar{X}}$, for each pair from each listener is given in Table 3-10. Notice that the number of responses per pair (N) and the number of practice judgments per pair are quite variable across hubs and listeners. The reliable within listener difference in magnitude relevant to single-factor order independence (E2) are displayed in Figure 3.5-3a,b. Orders are shown as directional arcs between lines radiating from a hub to two stimuli.

(a) The first part of Experiment 2b offered each listener fifteen tests of order independence (E2) in F1 and nine tests in F2. In F1, LL violated (E2) in three tests:

$$\text{LL } (\underline{27,212}) > (\underline{27,312}) \quad :: \quad \begin{cases} (27,232) < (27,332) \\ (26,232) < (26,332) \\ (26,212) < (26,312) \end{cases}$$

strongly confirmed (E2) in three tests and had nine "no-tests." A "no-test" is a test quadruple wherein one or both pairs has overlapping confidence intervals, thus producing a technical (or weak) confirmation.

In F₁, RB had six violations

$$\text{RB } \begin{cases} (27,212) > (27,312) \\ (27,222) > (27,322) \end{cases} \quad :: \quad \begin{cases} (27,232) < (27,332) \\ (26,232) < (26,332) \\ (26,212) < (26,312) \end{cases}$$

four strong confirmations, and five no-tests. BI had no no-tests, seven tests strongly confirming (E2), and eight violations in F₁

$$\text{BI } \begin{cases} (27,212) > (27,312) \\ (27,222) > (27,322) \end{cases} \quad :: \quad \begin{cases} (26,212) < (26,312) \\ (26,222) < (26,322) \\ (26,232) < (26,332) \\ (27,232) < (27,332) \end{cases}$$

Among the nine tests of decomposability in F₂, most of the quadruples in the listeners' data yielded no-tests. RB had seven no-tests, one confirmation, and one violation; BI had eight no-tests

and one confirmation. LL's data, in this case, were less noisy. She yielded two no-tests, two confirmations, and five violations in F_2 decomposability:

$$27,332 > 27,312 \quad :: \quad 27,232 < 27,212$$

$$\begin{array}{l} 28,222 > 28,212 \\ 27,322 > 27,312 \end{array} \quad :: \quad \begin{array}{l} 28,322 < 28,312 \\ 27,222 < 27,212 \end{array}$$

LL responded as if 312 and 27 were close together and far from the other stimuli. Both BI and RB responded as if both 312 and 322 were 'too close' to hub 27 and 'too far' from the other stimuli. Level data for the stimuli in this first part of Experiment 2b are given in Table 3-1 and in section 3.5.5. The listeners' judgments are certainly consistent with a hypothesis that L or L_2 is a salient auditory attribute in some circumstances. It seems, on inspection that a decomposable combination rule on F_1, F_2, L_2 with psychological scales monotonically related to the physical scales would easily fit the data in 2b. This issue and the problem of estimating a common perceptual unit for levels and frequencies will be taken up in section 3.8.

(b) The second part of Experiment 2b yielded two tests of F_2 order independence and two tests of F_3 order independence per listener. Among the three listeners, all the test quadruples except one yielded no-tests. That one test was LL: $1,5 > 1,9 \quad :: \quad 1,2 < 1,4$ which is an F_2 violation of the kind sought in this part of Experiment 2b and the kind expected from concomitant changes in L_2 . In interval notation,

$$I_2(1150,1300), I_3(1500,1700) > I_2(1000,1150), I_3(1500,1700) \quad ::$$

$$I_2(1150,1300), I_3(1700,2100) < I_2(1000,1150), I_3(1700,2100).$$

Among the eleven no-tests from the three listeners, if the means alone were accepted as order indicators, there would be five more violations, four of which depend on stimulus 5 or stimulus 3 being too far; one has stimulus 3 too close to the hub (see fig. 3.5-3b).

(c) 2b Summary. Violations of decomposability were successfully constructed with stimulus pairs that included F_1 or F_2 intervals that could be compared to like intervals near the $F_1 = F_2$ plane. In the first part of Experiment 2b, across listeners, 23 out of 41 strong

tests were in violation of the predictions of from decomposability in F_1, F_2, F_3 . In the second part of Experiment 2b, similar constructions near $F_1 = F_2$ were less successful.

3.5.7 Summary of Experiments 2a and 2b.

In the light of results from 2b, Experiment 2a showed that a systematic divergence from additivity, and even from decomposability, can go virtually unnoticed in a large scale experimental design, if stimuli are not chosen carefully with reference to an alternate representation. In Experiment 2a, among a thousand or more strong tests of interdimensional additivity, fewer than fifty tests violated (E3) - joint factor order independence. Using the results of Experiments 1 and 2a as pilot material in order to better place the hub stimuli, Experiment 2b produced violations in over 50% of the strong tests to be found in the data.

There are difficulties remaining in Experiment 2. F-patterns like $31x = (700, 800, x)$ or like $111 = (175, 800, 1800)$ are not common sounds often used as phonemic norms for vowels. Further, although Experiment 2b fairly surely shows L or L_2 to be perceptually salient and capable of causing violations of formant frequency decomposability, the results from Experiments 1 and 2a may indicate that these effects are tertiary.

3.6 Experiment 3

Experiment 3 attempted to identify more definitely the causes for the failure of interdimensional decomposability in F_1, F_2, F_3 . Violations, particularly in Experiment 2b, were sought and found near the $F_1 = F_2$ plane, and were presumed to be caused by the perceptual effects of L or L_2 .

Experiment 3 addressed two aspects of the hypothesized effects of the amplitude levels L and L_2 . (1) Let us assume that the auditory effect of different L intervals or L_2 intervals is the sole cause of incongruent distance orders. Then, configurations of stimuli (in any region of the formant space) that violate decomposable ordering should

fail to produce violations, when the stimuli involved are re-synthesized with constant L and constant L_2 . (2) Since in Experiments 1 and 2 the supposed effects of formant level on perception seem most obviously to intrude on the listeners' judgments when two formants are very close in frequency, perhaps it is not the extreme concomitant changes in level that cause the perceptual 'anomalies', but, rather, some other effect associated with formants in close proximity. For instance, suppose that the auditory system located spectral peaks for comparison in frequency, as is presumed in the equations of sections 2.2.3 and in equations (3a) through (3d). If two formants get so close in frequency to each other that there are no longer two separate peaks that the ear can discern, listeners might change their mode of coding; and, in turn, this could produce the incongruous distances to these stimuli. In other words, in all the cases where we have found violations of decomposability, there may be a confounding "formant fusion within a critical band" as well as, or instead of, our often cited pattern of inconsistent amplitude intervals.

If the perception of amplitude levels is the sole cause of the inconsistent distance orders, we should be able to remove all the inconsistencies by fixing all formant levels. This should be true, even though frequency dependent auditory effects (like "peak fusion") remain.

3.6.1 Method

Listeners. LL, BI, RB, JH.

Stimuli. Two sets of stimuli were used in Experiment 3. Both sets consisted of the same 13 F-patterns, but one set was produced on a serial formant synthesizer, KLATT, diagrammed in Figure 3.3-1a, while the other set was produced on a parallel formant synthesizer, DICHOT, diagrammed in Figure 3.3-1b. The important difference between these two methods of synthesis, serial and parallel, is that in serial synthesis the amplitude levels of the transfer function spectrum are automatically constrained to follow from the formant frequencies as in

Equations 1.1. However, parallel synthesis allows independent control of formant frequencies and formant levels. In the set of 'parallel' stimuli used in Experiment 3, the relative formant levels, L_n , were held constant across all the stimuli. All the parallel stimuli had relative formant levels appropriate for the F-pattern 500,1500,2500,3500. Thus, normalized to our previous decibel reference for L_n , the parallel sounds in Experiment 3 all had $L_1 = 43$, $L_2 = 30$, $L_3 = 22$, and $L_4 = 11$. Both sets of stimuli are shown in Figure 3.6-1, and their formant frequencies and formant levels are given in Table 3-11. All the stimuli lie in the $F_3 = 2500$ plane and their common source spectrum and temporal pattern are as in Figure 3.3-2, except that the time between the two stimuli in a pair was 200 msec. We may distinguish the parallel and serial sounds by a prefixed P or S, so for instance there are two hubs, S52 and P52.

Procedure. Within each set of stimuli, there were twelve pairs formed from the hub, 52, in the A position paired with each of the 12 other stimuli in the B position. Two orders of each set of 12 pairs were included in one 25-pair sub-block of a 100-pair session. The other three sub-blocks in the sessions were comprised of similar pairs - either parallel or serial - including other hubs from a larger design that was not analyzed. In general, the apparatus and procedure follows that described in sections 3.3.2 and 3.3.3. A total of 24 different pairs were analyzed as judged by four listeners. The number of practice trials and tabulated responses for the listeners is given in Table 3-12. The data were analyzed and magnitude ordering between pairs was derived as described in section 3.3.4.

3.6.4 Design

The original design of Experiment 3 included seven hubs and 21 non-hub stimuli, but only one hub, 52, and 12 other stimuli were selected for analysis. As can be seen from Figure 3.6-1, Experiment 3 developed distances with constituents I2(1000,1025) or I2(1025,1200) crossed with each of six F_1 intervals: (525,300), (525,400), (525,500),

(525,600), (525,700), and (525,800). Two of these F_1 intervals, $I_1(525,400)$ and $I_1(525,600)$ were also combined with $I_2(800,1025)$. All the axiomatic tests in the design relate to decomposability as reflected in single factor order independence (E2).

Among the serial sounds, violations were expected to involve stimuli S21, S20, and S15 being judged very dissimilar to the hub, S52. Consider the stimuli of Experiment 3 as represented in Figure 3.6-1. The stimulus points are arranged on the F_1 by F_2 plane in the same manner, whether produced by serial synthesis on the left or parallel synthesis on the right. For the parallel stimulus set, L_2 is constant at 30. For the serial stimuli, iso- L_2 contours describe the change in L_2 caused by changes in F_1 and F_2 . The L_2 contours are spaced at every 5 dB. Serial stimuli are constrained to lie on this curved L_2 surface in F_1, F_2, L_2 space.

Now consider the perceptual distance between the hub, S52, and stimuli 21 and 26. The distance in the F_1 by F_2 plane $d(P52, P26)$ is greater than $d(P52, P21)$. This assumes no contribution by differences in L_2 , as are necessarily found in the serial sounds. On the left side of Figure 3.6-1, notice that we must cross three (5 dB) level contours between S52 and S21, but only two contours between S52 and S26. So, in the serial case, we are faced with the opposition of a greater frequency interval paired with a smaller level interval, against a smaller frequency interval paired with a larger level interval. What will a listener do? Experiment 3 was designed to furnish the answer.

Assuming decomposability in F_1, F_2, L_2 , if the magnitude difference between $I_2(34,47)$ and $I_2(34,42)$ is larger than the magnitude of $I_2(1000,1200)$, then, for the serial sounds, listeners will rate $d(S52, S21) > d(S52, S26)$. Assuming $d(52,18)$ less than $d(52,23)$ for both sets of sounds, this violates (E2). Among the parallel sounds, no violation of (E2) would be expected, if vocoids are perceived as a decomposable combination of F_1, F_2 . This expected pattern would form a 'level-dependent' violation of decomposability in formant space for the serial sounds, but not for the parallel sounds.

TABLE 3-11: STIMULI OF EXPERIMENT 3.

Bandwidths: $B_1 = 50$, $B_2 = 70$, $B_3 = 110$ Serial

	Name	F_1	F_2	F_3	L_1	L_2	L_3
(hub)	52	525	1025	2500	45	34	12
	13	400	800	2500	42	31	2
	15	600	800	2500	50	43	9
	17	400	1000	2500	42	28	6
	18	500	1000	2500	45	33	10
	19	600	1000	2500	47	37	14
	20	700	1000	2500	49	42	17
	21	800	1000	2500	53	47	19
	22	400	1200	2500	41	26	10
	23	500	1200	2500	44	31	14
	24	600	1200	2500	45	35	18
	25	700	1200	2500	47	39	21
	26	800	1200	2500	49	42	23

Parallel

Names	F -patterns	L_1	L_2	L_3
as above;	as above			
may be		43	30	22
distinguished		for all 'P' stimuli		
as P13 vs S13				

TABLE 3-12: DATA FROM EXPERIMENT 3.

Listener Practice n N	<u>Serial</u>							
	LL (6) 14		BI (6) 12		RB (4) 10		JH (10) 12	
	$\bar{X}$	$2\sigma_{\bar{X}}$	$\bar{X}$	$2\sigma_{\bar{X}}$	$\bar{X}$	$2\sigma_{\bar{X}}$	$\bar{X}$	$2\sigma_{\bar{X}}$
Stimuli								
52,13	7.85	0.35	3.83	1.28	7.10	1.30	5.41	0.49
52,15	8.43	1.29	5.50	0.95	6.40	1.28	5.83	2.15
52,17	5.21	1.01	2.75	1.09	6.30	1.55	3.83	1.46
52,18	1.21	0.41	2.17	0.89	3.30	0.78	0.92	0.49
52,19	3.14	1.40	4.83	1.34	4.00	0.77	2.50	1.32
52,20	8.14	1.12	5.83	0.98	7.70	0.64	4.08	1.44
52,21	10.79	0.86	7.17	0.68	8.40	0.91	8.42	0.95
52,22	5.21	0.67	3.75	1.29	7.60	0.66	4.58	1.60
52,23	3.86	1.45	2.42	1.03	7.40	0.91	2.50	0.76
52,24	4.79	1.37	4.00	1.52	6.50	1.28	2.25	1.09
52,25	6.43	1.05	5.08	1.18	7.20	0.98	3.67	1.43
52,26	9.86	0.63	6.92	0.64	7.00	1.18	7.33	0.74

Listener Practice n N	<u>Parallel</u>							
	LL (6) 12		BI (6) 10		RB (2) 10		JH (6) 12	
	$\bar{X}$	$2\sigma_{\bar{X}}$	$\bar{X}$	$2\sigma_{\bar{X}}$	$\bar{X}$	$2\sigma_{\bar{X}}$	$\bar{X}$	$2\sigma_{\bar{X}}$
Stimuli								
52,13	7.75	1.29	4.30	1.18	7.50	1.11	5.33	0.74
52,15	5.08	1.89	3.40	1.28	5.90	0.83	3.67	1.31
52,17	5.33	1.37	3.20	0.98	7.40	0.91	3.41	0.64
52,18	1.58	0.64	1.60	1.11	3.30	0.45	0.75	0.43
52,19	3.50	1.65	2.80	1.88	4.10	0.83	2.17	0.89
52,20	8.33	1.37	4.90	1.13	7.50	0.92	4.58	1.49
52,21	9.67	1.74	6.30	1.10	7.20	1.16	7.08	1.75
52,22	6.75	1.09	4.60	1.28	8.20	0.60	5.17	1.72
52,23	4.92	1.32	2.30	1.27	7.80	0.87	1.42	0.75
52,24	4.83	1.72	3.10	0.94	4.80	1.07	2.00	0.91
52,25	7.50	1.55	4.50	1.20	6.90	1.04	3.58	1.18
52,26	9.67	1.31	6.50	0.80	7.20	1.24	6.67	1.49

3.6.3 Results

The mean response magnitudes, $\bar{X}$, to the pairs for each of the four listeners are given in Table 3-12. The magnitude that is added to and subtracted from the sample mean to form confidence intervals, $2\sigma_{\bar{X}}$, is also given (see sec. 3.3.4). For each listener, in relation to each stimulus set there were eight potential tests of order independence in F_1 , and eleven potential tests in F_2 . There were no violations of F_1 decomposability in any listener's data for either the set of serial sounds or the set of parallel sounds. In regard to F_2 decomposability, there was a small but regular pattern of violations in the data from three listeners when judging the serial sounds, but there were no F_2 violations for the parallel sounds.

Skipping the F_1 tests, the eleven potential tests of F_2 decomposability may be divided into "confirmations", "no-tests", and "violations" as in Experiment 2a (see Table 3-7a). In BI's data, all eleven of the distance quadruples that fit schema (E2) were no-tests in both the serial and the parallel data.

In the serial data from LL, there were four no-tests, three confirms, and four violations:

$$\begin{array}{l} 52,21 > 52,26 \\ 52,20 > 52,25 \end{array} \quad :: \quad \begin{array}{l} 52,19 < 52,24 \\ 52,18 < 52,23 \end{array}$$

LL's data for the parallel sounds yielded 11 no-tests in F_2 .

Listener RB produced one confirm, and eight no-tests in F_2 . The violations of F_2 order independence among the serial sounds were

$$52,21 > 52,26 \quad :: \quad \begin{array}{l} 52,19 < 52,24 \\ 52,18 < 52,23 \end{array}$$

For the parallel sounds, RB gave eleven no-tests.

JH produced a similar pattern on the serial stimuli. He had ten no-tests and one violation in F_2

$$52,21 > 52,26 \quad :: \quad 52,18 < 52,23.$$

Among the eleven potential tests of F_2 order independence in the parallel sounds, JH had three no-tests and six confirmations.

Summary. For two sets of stimuli, one with constant formant levels, listeners' data suggest that magnitude orders inconsistent with a decomposable model of auditory distance among vocoids are caused by auditory effects of formant amplitude level. The data in this experiment were only sufficient to show that F_2 intervals may be overridden by level changes. There were no violations of (E2) in F_1 , although opposing frequency and level intervals were in the design. However, three out of four listeners showed a consistent (and expected) pattern of violations of decomposability in F_2 within an F_1, F_2, F_3 representation of vocoid perception when stimuli had normal analytically determined concomitant formant level values. For stimuli with fixed formant levels, no judgments were incompatible with a decomposable formant frequency representation.

3.7 Experiment 4

Experiment 4 attempted to gather reliable data from which information about psychophysical scales for F_1 and F_2 could be inferred. The design of Experiment 4 included tests of two elemental ordinal conditions which must be satisfied for the construction of any additive numerical scale of formant frequency.

3.7.1 Method

Listeners. BI, LL, RB, and JH.

Stimuli. The constant properties of the stimuli in Experiment 4 are depicted in Figure 3.3-2, except for a difference in the presentation format noted under Procedure below. Five stimuli differing only in F_1 and five stimuli differing only in F_2 were synthesized. These ten stimuli are listed and described in Table 3-13 and shown in F_1 by F_2 in Figure 3.7-1. Notice that the F_1 stimuli are in the $F_3 = 2500$ plane and the F_2 stimuli are in $F_3 = 2800$.

Procedure. Experiment 4 consisted of one session of 100 pairs. Each sub-block of 25 pairs included two instances of each of the 10 pairs from one set of 5 stimuli yielding two quasi-random orders of the pairs

per sub-block. Responses to the first four pairs and the last pair within each sub-block were not recorded. The first two sub-blocks in the session presented the 10 pairs formed from the F1 stimuli illustrated in the right side of Figure 3.7-1. The last two sub-blocks in the session presented the 10 pairs formed from the F2 stimuli illustrated on the left side of Figure 3.7-1. Only the lower-frequency-first order was presented, e.g., 1,5 but not 5,1.

Because there was not a constant sound in the A position in the pairs, a 'timing cue' signal was presented before each pair in order to cue the listener. The timer signal sounded like a quiet 'twang' which preceded the first member of each pair by a second. This presentation format is shown in Figure 3.7-2 which can be compared to 3.3-2. The timing cue was produced by exciting a damped one pole resonance by a low intensity impulse train.

Data Analysis. Pairs were compared to strict inequality, ">", as in section 3.3.4.

3.7.2 Design

Experiment 4 provided tests of two axioms which in conjunction with certain technical assumptions are necessary to establish the compatibility of data with intradimensional subtractivity, (3b). These axioms apply to stimuli that only differ along a single stimulus dimension. Violation of the axioms indicates that no interval psychological scale can be constructed along that stimulus continuum. As developed by Tversky and Krantz (1970, p. 575-580), our Equation (3b) can be established from their axioms (A5) and (A6). We will formulate two ordinal relations (E4) and (E5) that are modifications of these two axioms, relying on the strong order, ">".

When applied to vocoid perception, Equation (3b) from section 3.2 says that real numbers may be assigned to various frequencies of FN such that the order of perceived similarities is the same as the order of magnitude differences among the real numbers. If we consider Equation (3b) as a hypothesis about the perception of vocoids, we can

define two basic ordinal predictions which may be called "between-ness transitivity" and "constant curvature". Before formulating them, however, we must introduce a definition of "between". We will say that "r is between s and p" if and only if $d(s,p) > d(s,r)$ and $d(s,p) > d(r,p)$. This "between" relation is symbolized as $s|r[p$. Notice that we have formulated "between" with strict inequalities, and that the between relation is symmetric in its first and third terms. Let us define "weak between-ness" as 'between' specified with " $\geq$ " and symbolized as $s|r|p$. The first ordinal prediction from intradimensional subtractivity (3b) is between-ness transitivity, (E4), which can be stated (following Tversky and Krantz):

$$\begin{aligned}
 &\text{with all other } F_m \text{ (} m \neq n \text{) fixed,} \\
 &\text{for any four values of } F_n; x, y, z, w, \\
 &\text{if } x|y|z, \\
 (E4) \quad &\text{then (i) if } y|z|w \text{ and } y \neq z, \text{ then } x|y|w \text{ and } x|z|w, \\
 &\text{and (ii) if } x|z|w, \text{ then } x|y|w \text{ and } y|z|w.
 \end{aligned}$$

I believe it can be shown that if weak sequential between-ness, WSB (equation 3.4b in sec. 3.4.3), is satisfied for a set of stimuli, then exactly the same 10 strong between relations must be satisfied for (E4) to hold. For either set of five stimuli in Experiment 4, there are ten instances, or tests, of weak sequential between-ness: $1|2|3$, $1|2|4$, $1|2|5$, $1|3|4$, $1|3|5$, $1|4|5$, $2|3|4$, $2|3|5$, $2|4|5$, $3|4|5$; and given these, the conjunction of the ten corresponding strong forms $(1]2[3, 1]2[4, \dots, 3]4[5)$ is equivalent to (E4).

The second prediction from intradimensional subtractivity is constant curvature, which is taken from Tversky and Krantz's axiom 6. Constant curvature along a single dimension can be formulated as follows:

$$\begin{aligned}
 &\text{with all other } F_m \text{ (} m \neq n \text{) fixed,} \\
 &\text{for any six values of } F_n; x, y, z, x', y', z', \\
 (E5) \quad &\text{if } x|y|z \text{ and } x'|y'|z' \text{ and } d(y,z) \leq d(y',z') \text{ and } d(x,y) < d(x',y'), \\
 &\text{then } d(x,z) < d(x',z').
 \end{aligned}$$

Among 5 stimuli varying on one dimension, there are six tests of (E5). Within either set of stimuli, x, y, z, x', y', z' may be 1,2,3,3,4,5 or 1,2,3,2,3,5 or 1,2,3,2,3,4 or 1,2,4,2,4,5 or 1,3,4,3,4,5 or 2,3,4,3,4,5.

If (E4) and (E5) are satisfied, then a limited auditory frequency scale for F1 or F2 may be constructed simply from the ordering of pairs by considering the ten pairs along F_1 or F_2 to represent four intervals (1-2, 2-3, 3-4, and 4-5) and the six adjacent concatenations of these intervals. For instance, 1,4 is the concatenation of 1-2, 2-3, and 3-4. In some circumstances, such a set of orders can determine a psychological scale up to a numerically specifiable ratio uniqueness (see Exercise 1, p. 34 in Krantz et al., 1970). Scaling directly from the judged magnitudes given by listeners in Experiment 4 will be considered in section 3.8.

3.7.3 Results

A sample of 20 or more responses from each listener to each of the 20 pairs was recorded; at least the first two sessions and the last session from each listener were discarded. The mean response, $\bar{X}$, and the sample standard deviation, s , for each listener and each stimulus pair are given in Table 3-14. In Table 3-15, the responses to pairs are arranged by formant and are ordered by $\bar{X}$, high to low. The strict ordering by non-overlap of confidence intervals (see sec. 3.3.4) is indicated by the curved lines in Table 3-15. An arc connecting two pairs means that the higher pair is strictly greater than the lower pair and strictly greater than all the pairs below the lower pair. Inspection of Table 3-15 reveals very close agreement among the listeners on the F1 pairs and general agreement about F2. JH seems more different from the others than they are among themselves.

We will now review the data in relation to predictions (E4) and (E5) from intradimensional subtractivity. First, the data from the F1 stimuli are considered, then the data from the F2 stimuli.

First Formant. The most efficient way to check for violations of (E4) when stimuli come from a physical continuum is to check for weak

TABLE 3-13: STIMULI FOR EXPERIMENT 4.

All Bandwidths: $B_1 = 50$, $B_2 = 70$, $B_3 = 110$

First formant scale

Name	F_1	F_2	F_3	L_1	L_2	L_3
1	200	1500	2500	34	13	4
2	400	1500	2500	41	26	16
3	600	1500	2500	44	34	23
4	700	1500	2500	45	37	26
5	800	1500	2500	47	40	29

Second formant scale

Name	F_1	F_2	F_3	L_1	L_2	L_3
1	500	1000	2800	45	33	12
2	500	1200	2800	44	30	16
3	500	1500	2800	43	29	21
4	500	1800	2800	43	30	26
5	500	2100	2800	43	33	31

TABLE 3-14: RESULTS FROM EXPERIMENT 4.

First formant

Listener N	LL 24		RB 24		BI 20		JH 28	
Stimulus	$\bar{X}$	S	$\bar{X}$	S	$\bar{X}$	S	$\bar{X}$	S
1,2	6.12	1.09	6.50	0.70	4.25	0.82	3.46	1.11
1,3	7.33	0.89	7.41	0.70	5.10	0.70	6.36	1.14
1,4	7.95	1.01	8.17	0.55	5.40	1.24	7.57	1.29
1,5	10.21	1.60	8.79	0.40	6.90	1.22	8.54	0.56
2,3	4.58	1.15	4.38	0.80	3.45	0.97	2.25	0.57
2,4	7.58	1.11	7.08	0.99	5.20	0.67	3.71	1.03
2,5	9.13	0.92	8.58	0.49	6.00	1.14	5.86	1.18
3,4	4.33	1.74	3.92	1.15	3.45	1.02	1.14	0.58
3,5	8.92	1.32	8.25	0.66	5.55	1.24	3.79	1.26
4,5	3.75	2.04	4.04	0.97	1.70	0.90	1.32	0.75

Second formant

Listener N	LL 24		RB 24		BI 20		JH 28	
Stimulus	$\bar{X}$	S	$\bar{X}$	S	$\bar{X}$	S	$\bar{X}$	S
1,2	4.21	1.04	3.63	0.80	1.35	0.65	1.04	0.86
1,3	5.83	0.84	6.92	0.70	3.55	0.92	3.25	1.37
1,4	7.58	1.18	7.67	0.62	4.75	0.69	4.75	1.32
1,5	9.54	1.22	8.75	0.52	5.95	1.11	7.68	1.03
2,3	3.54	0.99	6.25	1.16	2.70	0.71	1.29	0.52
2,4	6.88	1.16	7.71	0.73	5.00	0.70	3.32	1.22
2,5	9.79	1.11	8.67	0.47	5.80	0.81	6.89	0.93
3,4	4.13	1.83	4.42	1.35	4.30	1.05	1.25	0.68
3,5	9.38	1.03	8.25	0.66	6.10	0.76	5.60	1.26
4,5	7.75	1.58	5.00	1.15	3.00	0.94	2.89	1.14

TABLE 3-15: F1 AND F2 INTERVALS ORDERED BY MEANS. STRICT INEQUALITIES ARE INDICATED BY ARCS. THE INTERVAL AT THE TOP OF AN ARC IS STRICTLY GREATER THAN, ">", THE INTERVAL AT THE BOTTOM OF THE ARC AND ALL LOWER INTERVALS

First formant

RB

1,5
2,5
3,5
1,4
1,3
2,4
1,2
2,3
4,5
3,4

JH

1,5
1,4
1,3
2,5
3,5
2,4
1,2
2,3
4,5
3,4

BI

1,5
2,5
3,5
1,4
2,4
1,3
1,2
2,3 3,4
4,5

LL

1,5
2,5
3,5
1,4
2,4
1,3
1,2
2,3
3,4
4,5

Second formant

RB

1,5
2,5
3,5
2,4
1,4
1,3
2,3
4,5
3,4
1,2

JH

1,5
2,5
3,5
1,4
2,4
1,3
4,5
2,3
3,4
1,2

BI

3,5
1,5
2,5
2,4
1,4
3,4
1,3
4,5
2,3
1,2

LL

2,5
1,5
3,5
4,5
1,4
2,4
1,3
1,2
3,4
2,3

sequential between-ness, WSB (3.4b), and then check for strong sequential between-ness, SSB. While all the listeners' data from the F1 stimuli satisfied WSB, only JH showed 10 between relations. From among the data of the other three listeners, the following orders were needed for SSB, but were not found:

missing order	listener
$\sim(2,5 > 3,5)$	BI,LL,RB
$\sim(1,5 > 2,5)$	BI,RB
$\sim(1,4 > 2,4)$	BI,LL
$\sim(1,4 > 1,3)$	BI,LL
$\sim(2,5 > 2,4)$	BI

A few examples of weak between triples that violate (E4) in these listeners' data are;

BI: 5|1|2 and 1|2|3, but not (5|1|3),
 LL: 5|2|3 and 2|3|4, but not (5|2|4),
 RB: 5|1|2 and 1|2|4, but not (5|1|4).

There are many more examples of violations of (E4), but they are all derivable from the list of missing orders above. For instance, from the first missing order that is shared by these three listeners, $\sim(2,5 > 3,5)$, the relation 3|2|5 can be formed. Given 5|3|2 by WSB, the following shared violation of (E4) may be stated:

5|3|2 and 3|2|5, but not (5|2|5).

Among the six tests of (E5) in F1, LL and RB gave single strong violations. The other listeners, JH and BI, gave weak violations of constant curvature, as did LL and RB. The strong violations were

LL: 1,2 > 4,5 :: 1,4 < 2,5
 and RB: 1,2 > 3,4 and 2,3 > 4,5 :: 1,3 < 3,5.

Counting weak violations, RB had five tests out of six in violation of constant curvature, LL had four out of six violations, BI had three out of six tests in violation, and JH had two in six. Several of the violations seem to involve the interval 2-3, or 11(400,600) seeming larger alone and smaller when added to another interval.

Second formant. For the set of stimuli in F2 there were more violations of (E4) and relatively fewer violations of constant curvature, (E5).

As with the F1 stimuli, WSB was satisfied in the data from all listeners. Again, a subsequent check for SSB revealed that JH had the requisite ten strict between triples. Thus JH had no violations of (E4). However, in the data from the other three listeners there were more missing orders than there were in F1, and the missing orders were more consistent across listeners. The following orders needed for SSB were missing:

missing orders	listener
$\sim(1,5 > 2,5)$	BI,LL,RB
$\sim(2,5 > 3,5)$	BI,LL,RB
$\sim(1,4 > 2,4)$	BI,LL,RB
$\sim(1,5 > 3,5)$	BI,LL
$\sim(1,4 > 3,4)$	BI
$\sim(2,4 > 3,4)$	BI
$\sim(1,3 > 2,3)$	RB.

From these failures of SSB, many violations of (E4) can be formed, including the following examples which apply in common to BI, LL, and RB:

$5|1|2$ and $1|2|5$ but not $(5|1|5)$
 $5|1|2$ and $1|2|4$ but not $(5|1|4)$
 $4|1|2$ and $1|2|5$ but not $(4|1|5)$
 $5|2|3$ and $2|3|4$ but not $(5|2|4)$.

There are many more violations of (E4) that were produced in common by BI, LL, and RB, but they are all derivable from the missing orders for SSB. Suffice it to say that there are many more violations here than in F1 and the violations are more constant across listeners.

In F2 the six tests of constant curvature, (E5), turned up few violations. LL and JH had no violations of (E5), BI had only one weak violation, $(1,3 \geq 4,5 :: 1,4 < 3,5)$, while RB showed this same violation and another, both in strong form

RB: $1,3 > 4,5 :: 1,4 < 3,5$
 RB: $2,3 > 4,5 :: 2,4 < 3,5$.

With respect to constant curvature, F_2 seems to be more regular for most of the listeners than was F_1 .

3.7.3 Results Summary

In the results of Experiment 4, the data from JH are almost completely compatible with the hypothesis (3b) that an interval scale for formant frequency can be constructed. For both F_1 and F_2 , and in relation to both of the two major conditions for intra-dimensional subtractivity, JH showed no strong violations and it is likely that his two violations of (E5) in F_2 are due to our inexact experimental procedure.

In general, for both sets of stimuli, violations of elemental ordinal conditions (E4) and (E5) were found in the data from the other listeners. The F_1 continuum seemed to show some violations of (E5). The F_2 continuum elicited three seemingly reliable violations of (E4) from the three listeners BI, RB, and LL. There are sufficiently many violations of (E4) in F_2 and the violations are similar enough across these listeners to suggest that these violations would appear even with a finer psychoacoustic method. Notice further that an additive difference model of vocoid perception, like (3d), that included an appropriately scaled subtractive dimension on L_2 could fit the apparent pattern of between-ness intransitivity in F_2 . However, if L_2 were the only 'extra' dimension among logarithmically scaled formant frequency dimensions, an equivalence like $IL_2(29,33) \approx I_2(1000,2100)$ would be needed to cause the failure of E4, and such an equivalence does not seem likely.

Note finally that the F_2 stimuli are also unidimensional stimuli in $(F_2 - F_1)$ and F_2' . Our data in Experiment 4 indicate that it is impossible to scale F_2 into the real numbers as part of a decomposable structure on F_1, F_2, F_3 . If this turns out to be true, within decomposable models of perception in the dimension sets F_1, F_2' or $F_1, (F_2 - F_1)$, F_2' and $(F_2 - F_1)$ would be unscalable also.

3.8 Results Summary and Conclusion

The results of Experiments 1 through 4 indicate that all the major ordinal properties, (E1) - (E5), that are predicted from the simplest kinds of decomposable distance functions on (F_1, F_2, F_3) are systematically violated in our experimental data.

Our mode of testing was to consider equations (3a) - (3d) as hypotheses about the structure of dissimilarity among vocoids (see sec. 3.2). If predictions (E1) through (E5) from these hypotheses were consistently violated by listeners and if the violations themselves were predictable from some non-arbitrary alternative distance formula, then the violations could be called "systematic." The ability to construct sets of distances that are contrary to the predictions from a hypothesis is the best demonstration that the hypothesis is false.

The alternative hypothesis from which violations were constructed and around which we will organize the interpretation of the experimental results is based on Flanagan's (1955a,b; 1957a,b; 1965) findings, which indicate that both formant frequencies and formant levels are potentially audible. In this section, we will first review the most basic results from the four experiments reported (sec. 3.8.1), and then a few clear examples will be given to illustrate the kinds of data that lead to the rejection of a dimensional distance in the formant frequencies alone (sec. 3.8.2). A third subsection (3.8.3) will suggest some approaches to scale construction based on regression analysis. The last subsection of Chapter 3 proposes some distance formulas that may be used as $d_v(v_j, v_k)$ [see fig. 1.1-1].

3.8.1 Four schematic distance functions, (3a) - (3d), were presented on page 50. These four equations are models of the structure of similarities data. The equations imply certain elemental ordinal conditions that should be met in empirical data if the models are correct; (E2) is introduced on page 60 and (E3) on page 61; (E4) and (E5) are introduced on page 97.

In Experiment 1 (sec. 3.4), among 145 tests of (E3), there was only one violation and that appeared only in the data from one listener. The ordinal predictions from decomposability and interdimensional additivity were generally satisfied. However, a set of less rigorous (equality) predictions from decomposability were violated because constant frequency intervals seemed larger to the listeners when the intensity levels were greater.

The data from Experiment 2a (sec. 3.5) also generally confirmed the ordinal predictions of Equations (3a) and (3c). However, it was shown that even if the data of Experiments 1 and 2a had been produced by a perceptual process that operated on formant levels as dimensions and was not decomposable in (F_1, F_2, F_3) , very few violations of (E2) could have been expected because of the particular choice of stimuli. Experiment 2b repositioned several stimuli on the basis of information from Experiments 1 and 2a, and thus successfully constructed several very systematic violations of prediction (E2) from decomposability (3a). Data from each listener were contrary to the pattern of dissimilarities that would be expected if vocoid perception were dimensionally organized by formant frequencies alone. Furthermore, the data were quite consistent with the pattern that would be expected on the basis of a perceptual process based on both formant frequencies and formant levels.

Experiment 3 (sec. 3.6) attempted to verify that formant levels, and not some other property of the stimuli, caused the perceived similarities to be anomalous for a model of perception that is decomposable into formant frequency dimensions. As in Experiment 2b, a set of stimuli were selected in such a way that if listeners attended to both formant frequencies and formant levels, then their data would violate (F_1, F_2, F_3) decomposability. Every listener whose data yielded any strong tests of (E2) gave violations of "formant space" perception just as would be expected if formant levels were independently perceptible. A second set of stimuli, identical to the first set with regard to their positions in formant space, were synthesized

with fixed formant intensity levels. This second set of stimuli was presented to the same listeners and no violations of (F_1, F_2, F_3) decomposability were found. This strongly suggested that the incongruent dissimilarity orders were caused by changes in formant levels.

Experiment 4 (sec. 3.7) presented two sets of stimuli. In each set, stimuli varied along only one formant frequency, either F_1 or F_2 . Responses to the set of stimuli that varied on F_1 showed scattered violations of (E4) and (E5). For the stimuli that varied along F_2 , the responses from three out of four listeners were generally in agreement with (E5), but showed systematic violations of (E4). These violations, too, are consistent with the alternative hypothesis that formant levels are perceptually salient. L_2 is non-monotonically related to F_2 in the range of values used in Experiment 4. If L_2 intervals were scaled large enough relative to F_2 intervals and were incorporated into an interdimensionally additive model of perception on (F_1, F_2, L_2) , then this (F_1, F_2, L_2) model would predict the kind of violations of (E4) in (F_1, F_2, F_3) that were found in the data. One listener's data was in nearly perfect agreement with (E4) and (E5) for both the F_1 and the F_2 stimuli.

From these experiments there is firm evidence that the auditory perception of vocoid transfer functions is not representable as a decomposable structure with formant frequency dimensions. Several violations of single factor order independence (E2) were quite systematic. The violations were constructed from the alternative predictions of an alternative hypothesis and the violations were consistent across listeners. Furthermore, a fairly systematic failure of (E4) in the judgments of stimuli that varied along F_2 indicates that within a decomposable formant frequency representation of auditory distance, F_2 may not be scalable into the real numbers. If this failure of (E4) on F_2 is reliable, then dimensions that might replace F_2 , like $(F_2 - F_1)$, (F_2') , or $(F_2' - F_1)$, would also be unscalable within any decomposable structure where the remaining dimensions were scaled formant frequencies.

3.8.2 This subsection provides illustrative examples of the kinds of orders of auditory dissimilarity that violated the conditions for (or predictions from) distance formulas (3a) - (3d). It is hoped that these examples will clarify the meaning of the experimental data in relation to the usual geometric organization of vocoids in formant space.

Consider Figure 3.8-1a. There are ten points in a plane forming a 3x3 grid, with an extra point, J. By merely assuming the very weak structure (3a) on this set of points, we get the following several simple examples among many:

$$AC = DF = GI$$

$$DG = FI$$

$$AE > EC \text{ iff } DE > EF \text{ iff } GB > BI \dots$$

If normal Euclidean distance is assumed to work on the points in this F_1 by F_2 plane, then we get the above relations and more. The following are two simple examples:

$$AI = GC \text{ or } EI = FH$$

and

$$DF = DJ + JE + EF.$$

In contrast, our data typically show $AC < DF < GI$, although $AG \approx BH$. The data also show reliable order reversals between pairs of pairs like A,E : E,C and G,B : B,I. Results also indicate that the opposite diagonals of 'rectangular' figures were not always equal. For example, typically $EG > DH$, even though Euclidean distance would have them be equal. Lastly, Experiment 4 indicates that for most listeners

$$DF < DJ + JE + EF.$$

A close consideration of the position of these points, A - J, as they lie on an L_1 surface or an L_2 surface in Figure 3.8-1b,c suggests that most of the empirical relations that are anomalous in (F_1, F_2) plane geometry, could be handled in (F_1, F_2, L_1, L_2) solid

geometry. One tendency in the experimental data - $BB' < II'$ - cannot be fit into an additive distance by incorporating L_n dimensions.

3.8.3 One possible approach to scaling a set of perceptual dimensions would be to do a regression analysis directly on the listener's responses or on their mean magnitude responses. For instance, the data from Experiment 4 might be fit by regression, after the listeners were sorted into JH vs others.

For the data from JH, the judged magnitudes of the 10 intervals in each dimension could be considered as estimates of the slope at their midpoints. So, for instance, to estimate the frequency scale $\psi_2(f)$, JH's average response to I2(1000,1200) could be taken as an estimate of $\frac{\partial \psi(f)}{\partial f}$ at $f = 1100$. Integrating the regression lines from such an analysis would give interval scales in either formant frequency.

To best model the other subjects, the data for the F2 stimuli would be treated in a different manner. For instance, we may assume a restricted subset of dimensions and that a particular combination rule operates on them. In other words we might assume that the F2 stimuli are only heard to differ in F_2 and L_2 , and these dimensions combine in a Euclidean form. We would then have

$$d^2(v_j, v_k) = M(F_{2j} - F_{2k})^2 + (L_{2j} - L_{2k})^2$$

where M is the scale factor. It remains then to regress $(d^2(v_j, v_k) - (L_{2j} - L_{2k})^2)$ against $(M(F_{2j} - F_{2k})^2)$.

3.8.4 The experimental data will now be brought to bear on the justification of an overall dimensional distance formula for vocoid transfer functions. The step size used in Flanagan's quantization of $(F_1, F_2, F_3, L_1, L_2, L_3)$ may be taken as a point of departure (see p. 32). Flanagan's research suggests that dissimilarity among vocoids may be organized, roughly, as a six dimensional additive-subtractive distance scaled so that:

$$(3.8a) \quad I1(20) = I2(50) = I3(75) = IL1(1) = IL2(2) = IL3(5).$$

We may assume for the tests that satisfied (E2) and (E3), that some scales on F_1, F_2, F_3 are adequate to model the data. The relation $2F_1 = 5F_2 = 7.5F_3$ in (3.8a) is a tri-linear scale similar to $F_1 = 2F_2 = 3F_3$.

Let us examine the four most reliable violations of (E2) in (F_1, F_2, F_3) to see whether or not an additive representation on $(F_1, F_2, F_3, L_1, L_2, L_3)$ scaled as in (3.8a) will account for the orders found in the data. First, consider these F-patterns and their associated L-patterns.

name	F-pattern			L-pattern		
j	625	925	2300	48	41	13
k	625	1600	2300	44	36	27
m	350	800	3000	41	31	0
n	350	2000	3000	39	24	19
o	700	800	3000	55	54	11
p	700	2000	3000	43	37	32

In Experiment 2b, the following orders that seem anomalous in any usual scaling of decomposable (F_1, F_2, F_3) were found in the data from all three listeners: LL, BI, and RB;

- (a) $d(j, n) < d(j, p)$
- (b) $d(k, n) < d(k, p)$
- (c) $d(k, m) < d(k, o)$

From Experiment 3, the following dissimilarity ordering is anomalous with any usual scales in decomposable (F_1, F_2, F_3) (see Table 3-11 for formant levels):

- (d) $(525, 1025), (800, 1000) > (525, 1025), (800, 1200)$.

Among these four orders that violate decomposability (given $I_1(350, 625) > I_1(625, 700)$ and $I_2(1000, 1025) < I_2(1025, 1200)$), two are predicted by an additive structure in $(F_1, F_2, F_3, L_1, L_2, L_3)$ scaled as (3.8a) and two are anomalous even in the six dimensional additive representation.

In a six-dimensional additive representation with scales (3.8a), the pairs in (d) become:

$$\begin{aligned} \text{(left pair)} \quad & \frac{|525-800|}{20} + \frac{|1025-1000|}{50} + \frac{|2500-2500|}{75} + \frac{|45-53|}{1} \\ & + \frac{|34-47|}{2} + \frac{|12-19|}{5} \end{aligned}$$

$$\begin{aligned} \text{(right pair)} \quad & \frac{|525-800|}{20} + \frac{|1025-1200|}{50} + \frac{|2500-2500|}{75} + \frac{|45-49|}{1} \\ & + \frac{|34-42|}{2} + \frac{|12-23|}{5} \end{aligned}$$

or

$$\text{(d)} \quad 30.15 > 27.45,$$

which is the correct order. Likewise, the same distance gives the correct empirical order for (c):

$$\text{(c)} \quad 41.48 < 46.28.$$

However, an additive $(F_1, F_2, F_3, L_1, L_2, L_3)$ representation in linear scales does not account for orders (a) and (b). In fact, for both (a) and (b), an additive combination on the scales of (3.8a) strengthens the orders in the wrong direction.

The only subjective effect that has been discussed but is not included in (3.8a) is "augmentation". A given formant frequency interval is augmented when realized in formants of greater intensity. For instance, given two third formant frequencies 2000 Hz and 3000 Hz, the subjective difference between them depends on their levels - even when their levels are equal. That is, with all other F_n and L_n constant, $(F_3, L_3) = 2000, 25$ is much more different from $(F_3, L_3) = 3000, 25$ than $(2000, 10)$ is from $(3000, 10)$. This is similar to the augmentation effect reported for closely associated properties in vision. Differences in hue may appear greater when shown at a higher level of saturation. Krantz and Tversky (1975) found that the dissimilarity of visual shape was augmented when the (equal) area of the figures was increased. When a formant frequency interval is manifested in a

very soft, low power component of a sound, the frequency interval will seem smaller than it otherwise would. This augmentation effect is discussed under "equality predictions" from (3a) in Experiment 1.

One straightforward way to model an augmentation effect between two dimensions is by a polynomial combination rule. For each formant, FN, we propose two additive factors: (1) the absolute difference of the formant frequencies weighted by the formant levels, and (2) an absolute difference in formant levels. We propose, therefore, the following schematic auditory distance function for vocoids:

$$(3e) \quad d(v_j, v_k) = G \left[\sum_{i=1}^4 \left(|\psi_{ai}(L_{ij}) - \psi_{ai}(L_{ik})| + \phi_i(L_{ij}, L_{ik}) |\psi_{fi}(F_{ij}) - \psi_{fi}(F_{ik})| \right) \right]$$

Expression (3e) follows the notation conventions of (3a) - (3d) in section 3.2. Scaled appropriately, (3e) can fit the empirical orders (a), (b), (c), and (d) above. (3f) is an example of a scaling of (3e). (3f) follows the scales of (3.8a) fairly closely, for F-patterns in the neighborhood of (500,1500,2500), but is somewhat different even there.

$$(3f) \quad d(v_j, v_k) = \sum_{n=1}^3 \left[c_n (L_{nj}^{1.5} - L_{nk}^{1.5}) + c'_n \left(\frac{L_{nj} + L_{nk}}{60} \right) (|F_{nj} - F_{nk}|) \right]$$

with	$c_1 = .20$	$c'_1 = .05$
	$c_2 = .15$	$c'_2 = .025$
	$c_3 = .05$	$c'_3 = .01$

and L_n is normalized so that L_2 of (5000,1500,2500) is equal to 30.

(3f) is a hand-fit example of (3e). It yields the correct empirical orders for (c) and (d), but even more so than (3.8a). (3f) gives the pairs in (a) and (b) nearly equal magnitudes. A change which would increase the range of the function $\phi(L_{nj}, L_{nk})$ could yield the empirically correct orders on quadruples (a) and (b) above, but the results of Experiment 1 suggest that $\left(\frac{L_{nj} + L_{nk}}{60} \right)$ is justifiable.

Expression (3e) could be simplified by collapsing terms that are constrained to covary monotonically and by dropping out some less important terms. This sort of simplification could yield a distance like (3g).

$$(3g) \quad d(v_j, v_k) = G' \left[|\psi_a L_{2j} - \psi_a L_{2k}| + \phi(L_j, L_k) \left(\sum_{i=1}^3 |\psi_i F_{ij} - \psi_i F_{ik}| \right) \right]$$

Fitting schematic expression (3g) by hand to the data, a set of scales and coefficients is shown in (3h).

$$(3h) \quad d(v_j, v_k) = |L_{2j}^2 - L_{2k}^2| + \frac{L_j \cdot L_k}{1000} \left(\frac{|F_{1j}^2 - F_{1k}^2|}{1000} + \frac{|F_{2j} - F_{2k}|}{3} + \frac{|F_{3j} - F_{3k}|}{7} \right)$$

The decibel levels in (3h) are normalized so that in the F-pattern (500,1500,2500), L_2 is 30 and L is 80. Notice that either (3f) or (3h) will yield distances such that $DF < DJ + JE + EF$ in Figure 3.8-1.

Conclusion. There is a systematic failure of (3a). Expression (3e) is the main positive result of the psychoacoustic tests. (3e) takes into account the most noticeable types of divergence between the data and decomposable structure in F_1, F_2, F_3 . The major psychoacoustic qualities that are reflected in purely auditory judgments of dissimilarity can be fitted within a form like (3e). The coefficients and compromises of (3f), (3g), and (3h) are probably adequate, but certainly open to refinement.

CHAPTER FOUR

ARTICULATION AND DISTANCE

The first, and larger, part of this chapter describes the design and the input-output relations of a digital model of the vocal tract. The second, and shorter, part discusses a preliminary experiment designed to illuminate the relation between proprioceptive judgments of oral position and the physical positions of the tongue and jaw. Section 4.1 describes a less-than-state-of-the-art vocal tract model, presents its functional properties in graphic form, and evaluates their approximation to a human vocal tract. This vocal tract model will define the function $G(\vec{A}) = \vec{F}$. In section 4.2, an approximate inference will be made about the functions $F^{-1}(\vec{A}) = \vec{P}$ and $d_p(v_i, v_j)$, on the basis of the results of the proprioception experiment. These results, along with equation (3-f) from section 3.8, combine to complete Projects I and II, the determination of $H(G(F(\vec{P}))) = \vec{\Psi}$, and $d_p(v_i, v_j)$ and $d_\psi(v_i, v_j)$. (See fig. 1.1-1)

4.1 Digital Vocal Tract Model (VTM).

The organization of input to a vocal tract model (VTM) might be spatial or neural. The advantage of a neural organization is that information in speech must be coded within a neural framework at some stage, and such an input might better reflect the (language-independent) physiology of the vocal tract. For a discussion of these issues and a description of a physiologically oriented model of speech production, see Perkell (1974). The usual method in constructing VTMs is to manipulate the spatial position of gross anatomical parts of the supra-laryngeal vocal tract as seen in the mid-sagittal plane. A defense of the spatial approach to modeling articulation can be found in Lindblom (1972, p. 71-2). State-of-the-art VTMs are described by Mermelstein (1973) and by Lindblom and Sundberg (1971).

Consideration of the neuro-physiological structure of the supra-laryngeal vocal tract or just watching one's own tongue gymnastics in a mirror indicates that movements of the tongue and jaw can produce a great variety of resonator configurations. All of these shapes, if

vocoid, should be modeled here if results are to be completely phonetic. However, little is known about non-speech vocal postures, and most available data are from one of a few languages (English, Swedish, Russian, Japanese). The VTM constructed in this research program was designed to generate the mid-sagittal vocal tract outlines from the vocoids of one male American English speaker while he spoke English and English nonsense words (Perkell, 1969). The model could only generate outlines a little more extreme than this speaker's extreme vocoid postures from his stressed American English /i/, /æ/, /ɑ/, and /u/, and outlines in between these four over-extreme postures. Of course, the model's set of in between postures was constrained by the structure of the model, so the set of possible mid-sagittal outlines in the model is not necessarily just like those to be found for the original speaker.

Granted that this model is not satisfactorily language-independent, the evaluation of its appropriateness, even as an English-based VTM, must be in relation to the need for distance functions in working through the QNS and Maxi procedures (Project III). The VTM was evaluated in both its static and dynamic properties in the articulatory domain and the acoustic domain. (1) The VTM had to produce vocal tract outlines that resembled outlines from real vocoids and the VTM had to move from vocoid to vocoid along paths similar to those taken by real articulators. (2) In the acoustic domain, when the VTM outline closely resembled a given real outline, the F-pattern output of the VTM had to closely resemble the F-pattern produced by the real vocal tract. Furthermore, movement along some simple and realistic path in the VTM's articulatory space must result in a human-like path in formant space.

The general reader may want to skip to section 4.1.2.

4.1.1 VTM Design. The VTM worked in three parts, as shown in Figure 4.1-1. First, three input parameters determined a line which represented a mid-sagittal outline of the movable parts of the vocal tract in relation to the fixed surfaces (rear pharynx wall and roof of mouth). Second, the VTM calculated an area function of the vocal tract from this

mid-sagittal section giving the area of the vocal tract as a function of distance from the glottis. The third part of the VTM translated the area functions into acoustic transfer functions specified as formant frequencies, bandwidths, and phase. The first two parts of the VTM were programmed in FORTRAN on a PDP-9 computer at the M. I. T. Research Laboratory of Electronics. The first part was an original parametric design, fitted to the requirements of the second part which was simply an implementation of Maeda's (1972) method for converting mid-sagittal outlines into area functions. The third part was an existing subroutine available on the PDP-9.

a) The VTM calculated mid-sagittal outlines of the vocal tract on the basis of the six parameters: tongue x, tongue y, lip length, lip height, tongue root, and jaw height; although there were only three independent inputs - tongue x, tongue y, and 'rounding.' The six parameter structure was developed to integrate with Maeda's method of area function calculation.

Figure 4.1-2 shows the VTM in Cartesian coordinates and overlaid on an outline of the original vocal tract in [a] position. The stationary surfaces of the VTM are:

(4a) larynx front:	$y = -2x - 6.6$	$-13.5 \leq y \leq -11.0$
larynx back:	$x^2 + (y + 12.5)^2 = 4$	$-13.5 \leq y \leq -11.0, 1.3 < x$
rear pharynx wall:	$x = 0$	$-11.0 \leq y \leq -2.0$
velum-uvula:	$y = 2x/3 - 2$	$0.0 \leq x \leq 3.0$
palate:	$y = 0$	$3.0 \leq x \leq 7.5$
alveolar ridge:	$y = 7.5 - x$	$7.5 \leq x \leq 9.5$

with all constants in cms. The tongue body is a circle with radius = 3 cm, and with tongue x and tongue y the coordinates of the center of the circle. The limits of tongue x (TX) and tongue y (TY) are:

$$\begin{aligned} 3.3 < TX < 6.3 \\ -5.7 < TY < -3.2 \end{aligned}$$

The tongue root parameter and the jaw parameter in this VTM are completely defined by TX and TY in order to give a best linear fit to the concomitant measured jaw and tongue root positions in the original speaker. The jaw parameter was measured as the distance

between the upper and lower front teeth, and the tongue root was taken as the distance from a fixed point on the rear pharynx wall to the tongue. In the model, the tongue root parameter was the x coordinate of the tongue profile at $y = -8.0$. Tongue root (TR) and jaw (J) were calculated as follows:

$$TR = 3.7 + \left(\frac{4TY + 12.8}{5} \right) - \left(\frac{6.3 - TX}{6} \right)$$

$$J = .2 - \left(\frac{TY + 3.3}{3.0} \right) \cdot (1.0 + (.17 (TX - 3.3)))$$

These equations produced correct values of TR and J for /i/ and /a/, and were close for /æ/ and /u/.

In Figure 4.1-2a, the fixed structures in the VTM, as defined by equations (4a), are shown as wide heavy lines. The other side of the vocal tract outline in Figure 4.1-2a is movable and represents the tongue. The tongue outline is shown by a lighter line and is constructed as follows:

1. A straight line connects the top front of the larynx ($A = 2.2, -11.0$) to $(TR, -8.0)$.
2. Point $(TR, -8.0)$ is connected to the tongue body circle, $(x - TX) + (y - TY) = 9$, by a straight line tangent to the back of that circle.
3. A line parallel to the alveolar ridge is drawn from its tangent point on the upper front of the tongue body circle to its intersection with $y = x - 11.5$.

Notice that this last line represents a tongue blade fixed in a 'rest' position parallel to the alveolar ridge, thus precluding any retroflex vocoids.

The degree of lip rounding in the model was determined by a semi-parametric input variable, LIP, which took the values R, 1, 2, or 3, ranging from round to spread. For LIP equal to 1, 2, or 3, the lip height (LH) and lip length, (LL), are calculated as follows:

$$\text{intermediate variables: } YJ = 1.0 - \left(\frac{J - 0.2}{1.3} \right)$$

$$YL = \frac{LIP - 1}{2}$$

$$LH = J(1.0 + (YL + YJ)^2)$$

$$LL = 2.05 + 0.8J - 0.77LH$$

When LIP is equal to R, the procedure is the same as for LIP = 1 up to this point.

b) With a mid-sagittal outline derived from the three input parameters as above, the next step in the model was to convert that outline into a vocal tract length and define the cross sectional area (Cx) of the vocal tract tube at every distance from the glottis (Dx) to the outer end of the lips. The method that was used for obtaining the area function of the vocal tract was directly adapted from Maeda (1972), and had three parts: (1) for each point x along the midline of the vocal tract outline, calculate the sagittal dimension (Sx) and the distance along the midline from the glottis (Dx), (2) calculate the cross dimension (Cx) at each of these points and then (3) calculate the area (Ax) at these distances from the glottis. The equations for Cx, Ax, LW, and AL are taken from Meada (1972).

In part (1), after the midline was determined by simple geometric means, the distance, (Sx), from the tongue outline to the stationary outline was taken perpendicular to the midline at every 0.25 cm in the upper larynx, and above the tongue blade, and in the pharynx below TY. Along the regions bounded by the tongue body circle, sagittal distance, (Sx), was calculated along radii from (TX,TY) every $\pi/24$ radians, which is typically equal to about 0.5 cm along the midline. This yielded a set of ordered pairs (Sx,Dx). The distance from the glottis Dh, for the fixed point H (see fig. 4.1-2) can differ by as much as 2.5 cm when TX,TY = (3.3,-5.7) vs (3.5,-3.5). The lip section, which is attached just beyond fixed point H, also varies in length by a couple of cms or more depending on the values of LIP and (TX,TY).

Part (2) of the VTM calculated the cross dimension, C_x , at seven predetermined points along the vocal tract (G,B,C,D,E,F, and H in fig. 4.1-2) by the following formulæ:

$$\begin{aligned} C_g &= 4.49 - 0.14S_g - 0.38S_c - 0.26S_d \\ C_b &= 1.63 - 0.02S_b + 0.01S_c + 0.33S_d \\ C_c &= 1.95 - 0.33S_c + 0.54S_d \\ C_d &= 1.90 - 2.07S_d + 2.03S_c \\ C_e &= 1.07 + 0.79S_e + 1.05S_d \\ C_f &= 2.69(S_f)^{-0.66} - 1.75 \\ C_h &= 0.37 + 0.71(S_h)^{2.84} \end{aligned}$$

where S_x is the sagittal dimension at point X. A C_x was generated for all the S_x by linear interpolation between the fixed points D and H. Notice that in the upper pharyngeal region the C_x is inversely related to S_x . To insure that configurations were vocoid, any C_x less than 0.333 was set to 0.333.

(3) From the glottis (G) up to the midpoint of the velum (1.5, -1.0), the area of the vocal tract was modeled as an ellipse with S_x as one diameter and C_x as the other diameter. The area (A_x) for each x in this region was calculated as:

$$A_x = \pi/4 \cdot S_x \cdot C_x.$$

From the midpoint of the velum forward to the point H, the cross-sectional area of the vocal tract was modeled as the area between two parabolas, calculated as follows:

$$A_x = a_x \cdot S_x^{b_x} \cdot \left(\frac{S_x}{b_x + 1} - \frac{S_x - C_x}{1.5} \right)$$

with a_x and b_x defined at points E, F, and H as:

	a_x	b_x
E	1.4	0.5
F	1.1	0.5
H	1.4	0.2

and linearly interpolated in between and out to (1.5,-1.0). In the pharynx and upper larynx, the area A_x was only calculated at points G, B, C, and D and was linearly interpolated in between. From the back of the tongue body circle, D, to the teeth, H, the pairs (S_x, D_x)

were densely available and the other parameters, C_x , a_x , and b_x , were linearly interpolated for the area calculations.

The area of the lip section (AL) was calculated from lip height, LH, and lip width, LW, as an ellipse.

$$LW = 0.77 \cdot LH - 2.11 \cdot J + 2.5$$

$$AL = \pi/4 \cdot LH \cdot LW$$

When LIP was set to "R", the calculation proceeded with LIP = 1, except AL is finally set equal to AL/4.0. For LIP = R, LL is not changed from its value when LIP = 1.

The length of the vocal tract equaled $Dh + LL$ and areas were interpolated at every 0.1875 cm along the length, giving the final output of parts a) and b) as a straight tube comprising many small sections of varying cross-sectional areas.

c) Most of the last part of the VTM was accomplished by a sub-routine called TBFDA, developed by William Henke and available on the M. I. T. Speech Group computer. TBFDA computes the first four formant frequencies and bandwidths for tubes with non-constant cross-sectional area as a function of length. Tube area functions are approximated by a concatenation of smaller tube sections. The input parameters to TBFDA are: the number of tube sections; the length of each section; and a series of areas for the sections.

Because TBFDA models a tube with rigid walls, the lowest two formant frequencies were then corrected for the effect of soft walls by the formula (from Maeda)

$$F_n = \sqrt{F_n^2 + 165^2} .$$

4.1.2 VTM Output. In this section, a limit on the range of (TX,TY) will be established and the output of the VTM will be described and evaluated.

By allowing slightly more extreme tongue profiles than found in Perkell (1969), (TX,TY) equalling (6.3,-3.2) and (3.3,-5.7) were the positions for the center of the tongue body circle for [i]-like and

[α]-like gestures with the tongue profile nearly touching the stationary surfaces. However, (3.3,-3.2) produced a tongue outline that intersected the outline of the velum, and (6.3,-5.7) was much more extremely low and front than the real speaker's [æ]. (5.3,-4.7) was a bit more front than [æ], and (4.8,-5.2) was a bit lower than [æ]. From these positions, the low-front edge of possible tongue body movement was defined as

$$TY \geq TX - 10.0 .$$

The output of the VTM was sampled at 0.5 cm intervals in TX and TY at each of the four LIP values. Figure 4.1-3 gives F_1 , F_2 , and F_3 for all values of LIP (R,1,2,3) and (TX,TY) within the following region:

$$\begin{array}{ll} 3.3 \leq TX \leq 6.3 & TY \geq TX - 10.0 \\ -5.7 \leq TY \leq -3.2 & TY \leq 2TX/3 - 5.5. \end{array}$$

4.1.3 Evaluation. As suggested in section 4.1, above, the VTM had to resemble a real vocal tract in its static and dynamic properties in both the articulatory domain and the acoustic domain. To check the appropriateness of the VTM, two tests were made.

(1) Static Tests. A best approximation to the real speaker's articulation of [i], [æ], [α], and [u] was chosen from the range of (TX,TY) by superposition and visual inspection of the outlines from the glottis to the teeth. Even though the lip section was designed to simulate the mid-sagittal behavior of real lips in relation to their cross-sectional area at various (TX,TY), the lips can be considered as a lumped element (especially considering the calculations for LIP = R). So, at each (TX,TY) that best matched an outline for [i], [æ], or [α], an F-pattern at LIP = 2 or LIP = 3 (neutral or spread, respectively) was compared to the F-pattern measured from the real speech. For the (TX,TY) most similar to [u], an F-pattern from LIP = 1 or LIP = R (round or extra-round) was compared to the F-pattern of the real [u].

The best glottis-to-teeth outlines for matching [i], [æ], [ɑ], and [u] were found at the following values of (TX,TY) and are given with the associated F-patterns from the originals and from the VTM:

<u>vocoid</u>	<u>best (TX,TY) LIP</u>	<u>original F-pattern</u>	<u>VTM F-pattern</u>
[æ]	(4.8,-4.7) 3	650, 1650, 2575	586, 1742, 2800
[i]	(5.3,-3.3) 2	300, 2075, 2925	300, 1965, 2940
[ɑ]	(3.8,-5.2) 2	650, 1100, _____	602, 1355, 2240
[u]	(3.8,-3.3) 1	340, 1220, 2100	325, 1395, 1975

The superposition of (3.8,-5.2) on [ɑ] was shown in Figure 4.1-2b. The other (TX,TY)'s above are shown in Figure 4.1-4 superimposed on the original [æ], [i], and [u]. Notice that the match for [u] in Figure 4.1-4 is less satisfactory than the others. The poor match to /u/ is due, in part, to the difference in palate shape, but mostly to the difference in tongue body diameter found in the natural [u] which might be better fit with a tongue body circle with radius = 2.4 cm.

Best matches for acoustic output within $\sqrt{0.5}$ cm of the best articulatory (TX,TY) matches given above are:

<u>vocoid</u>	<u>(TX,TY) LIP</u>	<u>original F-pattern</u>	<u>Best VTM F-pattern</u>
[æ]	(4.8,-5.2) 3	650, 1650, 2575	650, 1646, 2838
[ɑ]	(3.3,-5.7) 2	650, 1100,	611, 1130, 2315
[u]	(3.3,-3.7) 1	340, 1220, 2100	396, 1220, 2050

From the articulatory and acoustic matches given, it seems that the VTM can come close to reproducing static articulatory-acoustic correspondences measured from the real speaker. Figure 4.1-5 shows the range of (F_1 , F_2) output from the VTM at each value of LIP. The dashed line approximately encloses the full range of the VTM if LIP were continuously varied. The VTM range is superimposed on the approximate (F_1 , F_2) range of a human adult male vocal tract (see section 3.1). The figure also shows the (F_1 , F_2) points for the modeled speaker's /i/, /æ/, /ɑ/, and /u/.

Two important inadequacies of the VTM's static behavior are 1) it cannot match [u] in the articulatory domain, and 2) the real speech on which it was based is English and does not reach good Cardinal [ɑ] or [u].

(2) Dynamic Tests. As measured by Kent (1972, p. 25-9), points on the tongue body of an (English) speaker move in nearly straight lines in the mid-sagittal plane with reference to the maxilla when the articulators are moving between vocoid targets. Linear mid-sagittal movement in relation to the maxilla is simple for our VTM, which leaves the question of whether or not linear movement in the VTM yields the same paths through formant space that are followed when a real (English speaking) vocal tract moves from one vocoid target to another. Two American English speakers (TB and JH from Chapter 3) were recorded while carefully pronouncing [iu], [ui], [iɑ], [ɑi], [ɑu], and [uɑ] in an a-syllabic manner (i.e., monotone, without epenthetic "y" or "w", neither clearly monosyllabic nor disyllabic). Both speakers also carefully pronounced "you", "we", "ya", "I", "how", and "wa" in order to check the 'a-syllabic' utterances. Figure 4.1-6 displays spectrographic pictures of "you" and [iu] as produced by JH, and shows the path of [iu] in an F_2 by F_3 plane. Later, the author, JB, recorded a-syllabic [æi], [iæ], [uæ], [æu], [æɑ], and [ɑæ]. Because all and only the utterances including [i] had consistent patterns of non-linear movement in formant space (reflecting an approach to $F_2 = F_3$ as one vocal tract resonance 'crossed' another), natural [iu], [ui], [iɑ], [ɑi], [iæ], and [æi] were chosen for comparison to the VTM's transitions between those same pairs. These resonance patterns may be compared to the patterns predictable from Fant's (1976) cavity-affiliation data.

The distances between LIP = R, 1, 2, and 3 were treated as 0.5 cm and the VTM followed identical paths between symmetric pairs. Figure 4.1-7 compares the VTM paths to the natural paths through formant space, and indicates that [iɑ] and [iu] are quite acceptable while [æi], although curved in the right direction, is less acceptable.

4.1.3 VTM Summary. A simple model was programmed to simulate vocoid vocal tract postures in the mid-sagittal plane, given a two dimensional tongue position input. From these postures and LIP values, area functions were calculated and translated into F-patterns, thus defining $G(\vec{A}) = \vec{F}$. The model's configuration and output at selected static positions was compared to the shape and output of the human vocal tract that it had been modeled on, and the VTM's dynamic output was compared to transitions between vocoids produced by other American English speakers.

4.2 Oral Proprioception

Although oral stereognosis has been studied, proprioception of articulator position has not been investigated. The goal of this study was to collect information helpful in identifying the one-to-one function $H^{-1}(\vec{A}) = \vec{P}$ mapping from an articulatory vector $(a_1, a_2, \dots, a_n)$ into its scaled counterpart $(p_1, p_2, \dots, p_m)$. The distance function was assumed to be Euclidean in form, since the articulators operate in physical space.

Because oral proprioception is an uncharted field, the goals of this experimental study were modest. For two relatively accessible articulatory continua (jaw elevation and tongue retraction), data were collected to indicate whether or not these continua could be manipulated in a manner linearly related unidimensional to physical distances. Although Project III requires a common unit in all articulatory parameters, no cross-parameter matches were collected. The experiment was preliminary and unrelated tests of linearity on two individual parameters were run. It is further recognized that subjects' ability to perform the experimental task is not necessarily related to the dynamic activities associated with language behavior.

4.2.1 Design and Procedure. A repeated physical measure of position associated with verbal labels was taken within subjects. The lowering of the jaw and retraction of the tongue were measured at their extremes (up-down, front-back) and were quartersected by asking for quasi-random

sequences of $1/4$, $1/2$, and $3/4$ (lowered or retracted) positions. Subjects were run one at a time and the results were tested for linearity.

Procedure. Subjects were instructed to try to move their tongues and jaws so that the resulting physical positions would be at five equally spaced points: the two extremes and the three quarter points. Subjects were told that the goal of the experiment was to see if people could accurately judge physical distance with the tongue and jaw, and without visual or verbal feedback. Subjects were asked to keep their tongue tips down and behind the lower front teeth. Subjects were instructed to tap a finger when they judged the articulator to be in the correct position (e.g., $1/2$ or $3/4$, as requested).

Each subject had a weighted thread glued to his tongue. The experiment did not proceed until the thread was clearly secure and the subject was accustomed to it. A plastic ball-point pen cap was positioned comfortably between the first molars on one side of the mouth to maintain a constant jaw position. A nose clip was fitted so that the tongue positions (especially extreme back) would be vocoid. After six practice moves to the extreme and $1/2$ positions, the test started. Fifty positions were announced in a quasi-random order, yielding 10 readings at each of the five positions. For the measures of jaw position, the thread was tied to one of the lower front teeth. The same sequence of positions was requested.

4.2.2 Subjects and Apparatus. Four male volunteers were tested. All were native speakers of American English without speech or hearing difficulties.

The study was conducted in a room with a high level of ambient broad band noise - allowing no auditory feedback for tongue position from differential resonance superimposed on the subjects' breathing. The subjects' heads were placed in a dental headrest and fixed by a band around the forehead. A fine non-elastic thread was attached to the tongue (approximately one inch from the tip when extended) or to the lower teeth. The thread was led through a screen and over a bar.

The thread was tensioned by the weight of a small washer which rode up and down with the movement of the articulator to which the thread was attached. On the opposite side of the screen, the experimenter read the elevation of the washer on a vertical meter stick. For measuring tongue movement, the bar was at the level of the mouth and about 10 cm from the subject's lips. When measuring jaw elevation, the bar was fixed at about 10-15 cm in front of the subject's forehead. (See fig. 4.2-1.)

4.2.3 Results. The results for the four subjects, for jaw and for tongue, are presented in Figure 4.2-2. The results are normalized by labeling the mean value of the tongue at "extreme front" and the jaw closed as "10.0 cm."

The following test of linearity was applied. If a straight line includes each confidence interval, then the data are considered linear; otherwise they are considered non-linear. In this restricted sense, the jaw data from RH and HG are considered linear. The other six graphs indicate non-linear tendencies. The non-linear data sets were primarily 'out of line' in that one or both of the extreme positions was placed too far away relative to the placement of the intermediate distances between $1/4$, $1/2$, and $3/4$ positions. The three intermediate positions $\{1/4, 1/2, 3/4\}$ passed the linearity test in all eight data sets.

Caution is required in interpreting these results. The extreme positions on the tongue retraction task were from 4.3 to 7.2 cm apart and were not common speech postures. The extreme front posture in 3 subjects typically had the tongue tip behind the lower front teeth. The dorsum of the tongue was touching the upper teeth ridge, so that the point where the thread was attached to the tongue was as far forward as the front surface of the front teeth, which forced lateral breathing. The extreme back positions, although more obscure, seemed equally peculiar. The jaw data seem less suspect, a priori, except that in very open positions the temporo-mandibular joint can pass from its normal range of predominantly hinge motion into a range of

hinge and sliding motion. The onset of the sliding action of the joint is far past the range of jaw movement commonly used in speech, but in the procedure of this experiment, the subject has to decide whether to define his extreme lowered jaw position before, at, or past the onset of sliding motion.

Several interpretations of the distance distortion in the data are possible: (1) less effort is required at the extreme postures (unlikely), or (2) control is less exact at or near the extremes, or (3) the extreme lower jaw position and the tongue extremes are irrelevant to the range of movement represented in the VTM. The last interpretation is accepted, thereby supporting the adoption of physical units in the VTM parameters. Euclidean $d_a(v_i, v_j)$, will substitute for the hoped-for $d_p(v_i, v_j)$. The function H will have to be ignored, limiting the extent of the phonetic channel as schematized in Figure 1.1-1 to $G(F(\vec{A})) = \vec{\Psi}$.

4.2.4 Summary. The experiment gathered no information about the relative magnitudes of intervals on different articulatory parameters, nor about rules for combining intervals into a composite distance. The investigation did indicate that physical units of tongue retraction and jaw height may form appropriate scales for use in Project III. This lead to the adoption of $d_a(v_i, v_j)$, Euclidean distance on the VTM parameters, in Project III. We further assumed, somewhat arbitrarily, that LIP intervals (R,1), (1,2), and (2,3) may be taken as equivalent to a 0.5 or 1.0 cm tongue movement.

4.3 Conclusion.

In this chapter, the solution of Project II was aimed for. We sought a function from self-perceived articulatory position into formant frequencies, $G(H(\vec{P})) = \vec{F}$, and a proprioceptive distance function in articulation. A vocal tract model was described that established a mapping from three articulatory input parameters, $\vec{A} = (a_1, a_2, a_3)$, into the resulting first three formant frequencies, $\vec{F} = (f_1, f_2, f_3)$. This VTM defines $G(\vec{A}) = \vec{F}$ as it will be used in Project III. A

preliminary study of the relation between physical measures of articulatory position and proprioceptive estimates of articulatory position was inconclusive, so establishment of a function H from $\vec{A}$ to $\vec{P}$ was abandoned and a Euclidean metric on the VTM inputs, $\vec{A}$, was accepted. The final outcome of Projects I and II, therefore, is a function $F(G(\vec{A})) = \vec{\Psi}$ and two distances, $d_a(v_i, v_j)$ and $d_\psi(v_i, v_j)$.

CHAPTER FIVE APPLICATIONS

Chapter 5 reports the application of the results of Chapters 3 and 4 to predicting vowel inventories. The vocal tract model described in Chapter 4 and the inter-vocoid distance formula developed in Chapter 3 may be taken together as a model of the transmission of vocoids from an articulatory representation, $\vec{A}$, to an auditory representation, $\vec{\Psi}$ (see fig. 1.1-1). The first part of this chapter (sec. 5.1) reports the design of a computer program to locate maximally distinct points in (F_1, F_2) . The second part (sec. 5.2) describes a procedure for mapping the gradience of our vocal tract model in relation to auditory quality.

5.1 Implementing Maxi

We ran the initial part of Maxi, wherein n vowels are positioned in a configuration that maximizes aggregate distance. This implementation is reasonably reliable because the shape of the (F_1, F_2) area that can be reached by a single human vocal tract is known and the form of our auditory distance formula (3e) is fairly well grounded. Below we will first describe a computer program that calculated a sample of distances in (F_1, F_2) . Then the output of the program will be discussed.

5.1.1. Maxi Program. The area of possible (F_1, F_2) was formed by an approximate quadrilateral with three rounded corners. The vertices of the quadrilateral were (250,2400), (250,600), (700,900), and (700,1700). This 'vocoid space' was sampled at 51 points, approximately every 200 Hz in F_2 and at every 100 Hz in F_1 . The 51 points are shown in Figure 5.1-1. Figure 5.1-1 also shows contours of constant L_1 and constant L_2 at intervals of 5 dB. F_3 was fixed at 2800 Hz.

For predicting the 'most favored' positions for a set of n vowels, n "marks" or "pointers" move around vocoid space. We ran

the program with $n = 3, 4$, and 5 . Because the 'true' perceptual position of $[i]$ is somewhere beyond (250,2400), one "mark" was fixed at that vertex. Two, three, or four other marks were then moved around in the vocoid space so that an aggregate distance among the n marks was calculated for each possible position of the moveable marks in the remaining 50 sample points. For each configuration of points, the inter-point distances were calculated and summed. The distance formula used in the program was:

$$d(v_i, v_j) = \sum_{n=1}^2 \left[\left(\left| \frac{L_{ni}^{15} - L_{nj}^{15}}{10} \right| \right) + c_n \left(\frac{L_{ni} + L_{nj}}{60} \right)^{1/2} \cdot |F_{ni} - F_{nj}| \right]$$

$$c_1 = .05 \qquad c_2 = .03$$

Notice that the range of the multiplying factor has been reduced by taking the square root. L_1 and L_2 have the same scale. Beside that, this distance is (3f) reduced to two dimensions.

5.1.2. Results. The maximum three-, four-, and five-point configurations are pictured in Figure 5.1-2a,b,c. I believe that these configurations are also the maximum configurations when all points are allowed to move. Arrows in Figure 5.1-2a,b,c indicate the direction and amount of shift that each of the points could move without reducing the aggregate distance by more than 1%. For instance, fixing the other three points in Figure 5.1-2b, if the higher of the two lower right points is moved to the end of its arrow, the aggregate distance will be reduced by less than 1%.

Figure 5.1-3 displays several different sets of three, four, and five points. Each of the constellations of points in Figure 5.1-3 has an aggregate distance that is within 5% of the corresponding optimum constellation shown in Figure 5.1-2. One further aspect of the output from our maximization program deserves note. As far as can be seen from the total computer output on these 51 sample points, there are local minima in these n -vocoid spaces.

The maximized configurations, especially for $n = 4$ and $n = 5$, are not as would be expected. This might suggest that the distance formula, or the vocoid space, or the Maxi principle itself is not correct.

5.2 Implementing QNS

Although our 3-parameter vocal tract model (VTM) is very limited when compared to the possibilities of human articulation, the model is an improvement over the older two- or four-section vocal tract models (see sec. 4.1.3). The output of the VTM was calculated at every 0.5 cm in its two tongue body coordinates TX and TY, for each value of LIP (R,1,2,3). The ranges of TX and TY are given on page 121. This sample of points in the articulation space forms a three-dimensional grid.

Since QNS posits a search for the positions of smallest gradience in the articulation space, the gradients must be ordered in some way. The gradient vector at each point in the grid of sampled articulations was approximated by calculating the average auditory distance between the sample point and each of its six co-axial neighbors in the grid. Thus the (approximate) gradience is reduced to a scalar quantity. This procedure was carried out twice; once by hand with a very simplified auditory distance, and once by computer with a modified version of Equation (3f).

5.2.1. Hand Implementation. Inputs to the by-hand running of the QNS procedure were the distance formula

$$d(v_j, v_k) = \left(\frac{F_{1j} - F_{1k}}{20} \right)^{1/2} \cdot \left(\frac{|F_{1j}^2 - F_{1k}^2|}{1000} + \frac{|F_{2j} - F_{2k}|}{3} + \frac{|F_{3j} - F_{3k}|}{5} \right)$$

and those regions of articulation in the model which had the most clearly acceptable and human-like acoustic behavior. These were the areas with high tongue position (front or back) and the areas with backed tongue positions (high or low). A gradient of auditory change

per unit articulatory movement was hand calculated only in these regions of the model. In the calculations it is assumed that lip-opening is linear in relation to lip height, with the full range of lip spreading and lip rounding at any one tongue position equivalent to the range of front-back movement of the tongue body in the model, i.e., 30 mm. Thus, in this hand calculated implementation, a difference of one step in the lip parameter, with fixed (TX,TY), is equal to 1.0 cm or two grid positions in TX or TY. Thus, the LIP sections shown in Figure 5.2-1 are effectively one centimeter apart.

Figure 5.2-1 displays the gradience of our VTM at the most clearly human-like area of (TX,TY) at LIP = R,1,2,3. The most stable region turned out to be a large area in the high-front tongue positions, both round and spread, which included the vocoids [i], [y], and [ɨ]. Another very stable region of articulation covered a portion of the mid-back-round area centering on the vocoid [ə] and continuing up into [o] and down into [ɔ]. The gradient was steep around [ɑ].

5.2.2. Computer Implementation. In this application of the QNS procedure the whole range of (TX,TY) was investigated. The distance equation used to approximate the gradience was very similar to (3f).

$$d(v_j, v_k) = \sum_{i=1}^3 \left[|L_{ij}^{1.5} - L_{ik}^{1.5}| + c_i \left(\frac{L_{ij} + L_{ik}}{60} \right)^{1/2} \cdot |F_{ij} - F_{ik}| \right]$$

$$\begin{array}{ll} c_1 = 0.2 & c'_1 = 0.05 \\ c_2 = 0.1 & c'_2 = 0.03 \\ c_3 = 0.05 & c'_3 = 0.01 \end{array}$$

Another difference in this application of the QNS procedure to Chapter 4's vocal tract model is that the distance between adjacent levels of the LIP variable was evaluated as 0.5 cm, as opposed to 1.0 cm in the hand-run version above. In effect, this emphasizes the relative contribution of lip articulation.

The output of this formulation of gradience is depicted in Figure 5.2-2. Figure 5.2-2 has the same format as Figure 4.1-3 and the

F-pattern of any point on Figure 5.2-2 can be read from Figure 4.1-3. The most stable articulatory position as shown in Figure 5.2-2 is at $(3.8, -5.2)_2$, or one point over from the best match to the original speaker's English /a/. This stable region around [a] is also nearly the largest such region in a (TX,TY) section. The next most stable position in the articulatory space is in the extreme upper right corner of each LIP section, with the smallest gradient at LIP = 2. This articulation is extreme high-front in phonetic terms; its mid-sagittal outline is far beyond the [i] posture of the original speaker. Note, however, that its F-pattern is not at or beyond Cardinal [i].

There is a third major region of relative auditory stability among high-central-round positions. This region is separate from the region in the upper-right corner as seen in the LIP = R and 2 sections. Notice that the most stable point in this region, $(5.3, -3.3)_2$, is exactly at the best articulatory and acoustic match to [i] as produced by the speaker on whom the model is based.

Conclusions. When the first part of the Maxi procedure is applied to a typical (F_1, F_2) vocoid area, the results are counter-intuitive for $n = 4$ or 5. This is probably because our intuitions have been schooled on traditional vowel charts. Three reasons for this 'failure' suggest themselves. (1) The distance formula used may be overly specific. A three or four percent difference will give us the 'correct' output. (2) The shape of the vowel space may be off. Perhaps if we unrounded the corner near (250,700), more maximal or near maximal configurations of points would settle there. Another problem with the "shape" of the space is its constant F_3 . Typically, when $F_1, F_2 = (250, 2400)$ $F_3 = 3000$, but when $F_1, F_2 = (250, 700)$ $F_3 = 2000$. As discussed in Chapter 3, such difference in F_3 can be easily heard under some conditions. (3) The last reason is that Maxi itself may be incorrect. Some possibilities in this vein are suggested in section 6.1.

Our implementation of QNS on Chapter 4's vocal tract model may be considered a bit more 'successful.' /i,a,u/ were expected, and broad stable regions were found at [i] and [a] in the fuller computerized application. Gradience was relatively steep in the high-back vocoid positions in both QNS runs. Remember (sec. 4.1.3) that notwithstanding an acceptable acoustic match to the modeled speaker's English /u/, the sagittal outline of /u/ was far off, and no sound at all near Cardinal [u] was possible in the VTM. A last note on QNS as applied to our VTM: with either distance and either scaling of LIP movement the steepest gradience at any sampled point is hardly more than three times as large as the most stable point. This ratio does not seem large enough to form the basis of vowel phonology.

CHAPTER SIX

CAVEATS AND CONCLUSIONS

Chapter 6 is divided into two sections. Section 6.1 is a warning (caveat) to the reader. Whatever the merits of the phonetic model developed in Chapters 3 and 4, the application of that model in Project III ignores many complexities of both phonetics and phonology. Section 6.2 is the general conclusion. Section 6.2 reviews several findings and suggests which aspects of Projects I and II are most critical to the outcome of Project III. This last concluding section also suggests which parts of our research most need further verification or refinement.

6.1 Caveats

All attempts at atomistic interpretation, which necessarily explain only one aspect or a single phenomenon and are therefore never comprehensive, are clearly inadequate. ... The appearance of single sounds must not be treated in an isolated fashion without regard for their place in the sound system.

Roman Jakobson (1968, p. 67)

This "caveats" section proposes to formulate some unresolved possibilities in phonetics and phonology in order to warn the reader. Stevens' theory of the Quantal Nature of Speech (QNS) and Lindblom's Maximization Principle (Maxi) were applied in Chapter 5 in a manner that might be characterized as "myopic." As applied, the QNS and Maxi hypotheses were very limited in scope. Vowel systems were developed as constellations of static points in a space of dimensions that only represented supra-laryngeal articulation.

In essence, the three subsections of this caveat section will posit three kinds of assumptions that might be made. The first assumption is that a larger range of sounds and more types of phonetic difference than were input to Project III could distinguish vowels within possible phonological systems. This most obvious limitation

of Project III is discussed in section 6.1.1. The second assumption (section 6.1.2) is that sound systems may be constrained by two or more extra-linguistic influences at once. It is possible that the 'Principle of Least Effort' operates in coordination with other human tendencies. The last "caveat" subsection (6.1.3) discusses two fundamental propositions that are assumed in the operation of both QNS and Maxi: human systems are optimal; and phonemes are identified by positive properties.

However, before we discuss the scope of Project III, two preliminary disclaimers are due.

Disclaimer one. Even as they were illustrating their principles of sound structure on limited sets of sounds, both Stevens and Lindblom knew that their procedures for explaining sound systems should properly be applied to whole sound systems and in reference to the entire potential of the phonetic channel. The fact that Project III focused on predicting just systems of *n* vowels is an important shortcoming of our research program. All the aspects of phonology discussed in sections 6.1.1, 6.1.2, and 6.1.3 should ideally be considered with reference to the simultaneous prediction (or development) of whole sound systems.

Disclaimer two. One should be wary of secondary and tertiary reports about the phonetics and phonemics of particular languages. Scholars interested in vowel-system typology often have to make phonetic inferences from published phonemic inventories that may reflect complicated analytic and/or typographic compromises. Stevens and Lindblom both cite /i/, /a/, and /u/ as nearly universal vowels and both unequivocally identify /a/ as a low-back sound with $F = (700, 1000)$. However, [a] - that most unmarked of unmarked vowels - is problematic right from the start. When Liljenkrants and Lindblom (1972, p. 845) give examples of 3-vowel systems in /i,a,u/, they cite Trubetzkoy's (1969, p. 106) examples which include Arabic and New Persian. They also cite

Hockett's (1955, p. 84) examples; including Iraqi Arabic. However, the more one examines the primary literature on Arabic and Persian the less satisfactory an /i,a,u/ description seems. Lambton (1971) identifies 10 (or perhaps even 12) different vowels in modern Persian - six (or eight) simples and four diphthongs. The major allophone of one of the Persian vowels is approximately [ɑ]. However, the most characteristic allophone of /a/ in Modern Standard Arabic (Abboud et al, 1968) and in Iraqi Arabic (Erwin, 1963) ranges within the American English vowels of "had", "tab", "bet", "bat", and "but". Arabic /a/ is more like [æ] than [ɑ]. These authors also credit Arabic with a system of six or seven monophthongal vowel qualities; [ɑ] being a variant of /a/ when adjacent to emphatics. This interlude on /a/ in Arabic and Persian has only been to caution the reader that assertions about particular vowel systems (or any phonological phenomenon) are not usually as simple as they appear at a distance. Occasionally, in the following three subsections we will refer to phonological 'facts' about particular languages. The reader should approach these statements with caution.

6.1.1. Caveat: All vowels are not simple vocoids distinguished by static superlaryngeal attributes.

The first problem with applying QNS and Maxi as in Chapter 5 is the severely limited range of phonetic parameters. For Maxi, F_1, F_2 patterns (F_1, F_2) were inputs and for QNS, very simple three-dimensional articulatory vectors $\vec{A} = (\text{hi, front, round})$ were input. However, may we not assume that other modes of manipulating the timbre of syllabic nuclei like nasalization, retroflexion, or pharyngealization should be included? Why should source parameters like F_0 , amplitude, or 'breathiness' not be included? Why should temporal properties like duration or dynamism not be considered potential means of distinguishing syllabic nuclei? Why not include at least some continuant contours as potential vowels? There are phonological systems that do use these phonetic properties to differentiate syllabic nuclei (see Jones, 1950; Sedlack, 1969; and Lehiste, 1970).

The limitations of the vocal tract model described in Chapter 4 are obvious. The model's articulatory range is based on only four examples of stressed phonemic vowels from one American English speaker. The model should properly have been based on the oral gymnastics of a skilled practical phonetician. The vocal tract model did not even approach the acoustic value of a good extreme Cardinal [i] or [u]. The tongue root, tongue tip and jaw are all controllable independent of tongue position and lip rounding, but were not allowed free movement in the model. The F-patterns appropriate to retroflex vocoid articulations were not included in the Maxi part of Project III; nor were formant level variations characteristic of nasal-vocoids used as inputs.

Concurrent distinctions.

We assume that people typically identify objects on the basis of information about several qualities of the object. Distinctions, like the tongue retraction in the articulatory domain or voice fundamental frequency, F_0 , in the acoustic domain, may be used as independent phonological features or they may vary redundantly with other features. An example of an independent phonetic parameter that was ignored in Project III is overall level, L. Consider / ϵ / in English 'próject' (noun) vs 'projéct' (verb). L is separately controlled in or below the larynx, and is used as a phonological feature. The greater overall level of /a/ in 'had' vs /i/ in 'heed', however, is usually considered to be a redundant property.

An important set of potentially independent phonological features has been overlooked in Project III; in articulatory terms they are laryngeal parameters and in acoustic terms they are source features. Phonetic and phonological features associated with the larynx are discussed by Catford (1964, 1968) and by Stevens and Halle (1971). Changes in F_0 , intensity, and 'voice quality' are fairly independent ways in which the larynx can change the perceptible quality of voiced syllabic nuclei. Certainly, a single supra-laryngeal sequence, like

[be], can be said with all eight possible combinations of higher or lower F_0 , greater or lesser source amplitude, and breathy or clear voice quality. F_0 is an extremely salient perceptual attribute of vocoids; where Flanagan (1957) quantizes F_1 in 14 steps, he suggests that F_0 be quantized into 40 steps for good quality vocoding. Catford (1968, p. 321) says that "probably the majority of languages in the world are tone-languages" in which significant F_0 differences are associated with shorter phonological units, like syllables. Differences in voice intensity are also quite easily produced and heard, but whereas tone systems of four static levels are not uncommon, systems of word stress rarely have more than two degrees (Jones, 1950). Different 'voice qualities' are also easy to hear, and Catford (1964) lists and analyzes many types that can potentially be used to differentiate vowels. The acoustic manifestation of voice quality (as separated from F_0 and intensity) is in the spectrum of the source to the vocal tract transfer function. As many as three or four distinctive voice qualities are regularly used in various languages, although usually there are no more than two different phonemic voice qualities in a language.

We may assume that another kind of potential phonological distinction works in time. For instance, two vowels may be distinguished by the lengths of time they are maintained. With vowels, across languages, ranging in duration from about 50 to 300 msec, and having a difference limen for duration in the range of 15% of duration (Lehiste, 1970, p. 12), one might imagine vowel systems built from a small number of vocoid F-patterns each at three or four distinctive lengths. Indeed, many languages use length to distinguish syllabic nuclei. However, phonetically long vowels are often phonemically geminate.

Dynamic-Static allows another kind of opposition that can happen in time. For almost any of the independent articulatory or acoustic phonetic parameters discussed previously, a vowel (or more generally a syllabic nucleus) may be phonemically distinguished by the essential

presence of change. F_0 can be used in a set of four level tones, (1-2-3-4), high to low, or in a group of four dynamic patterns in two levels, like (hi, low, hi-low, and lo-hi). Vowels may be monophthongal or diphthongal. It would be difficult to demonstrate why voice quality, intensity, and nasality, also, could not form distinctions in time. The vocoid nucleus of a syllable could start or end nasal in opposition to continuous nasalization, likewise intensity, or voice quality (for instance, breathy voice) could form oppositions of time course. Time course oppositions for tone (F_0) and vowel quality (F-pattern) are commonly used paradigmatically and syntagmatically in languages. Differences in dynamic patterns of nasality, voice quality, or intensity, are rarely used in such a way that they fit into a phonological system at the segmental level. Pike has suggested that voice quality and voice intensity are usually used in syntagmatic oppositions that range in time over several segments because certain articulatory structures have 'longer time constants' or 'greater inertia'.

All of the above concurrent distinctions have been assumed to be potential distinctive features. However, as best described by Lehiste (1970), and Jones (1950), phonetic distinctions are often not phonologically independent, but singly or even multiply redundant. The American English vowels, considered loosely, yield many examples of multiple phonetic redundancy. Take as an example the vowels in "bet" and "bait". Roughly, / ϵ / and /e/ in English differ by one phonological opposition which is variously identified as 'height' or 'length' or 'tensity'. The phonetic differences, however, are manifold (Lehiste, 1967, chapters 9, 16, 17, 20, 22). / ϵ / is a monophthong, /e/ $\approx$ [e^i] is a diphthong. The F-pattern of the nucleus of / ϵ / is (550,1650); /e/ is about (350,2050). /e/ is typically about 30 percent longer in duration than / ϵ /, about two or three decibels lower in overall amplitude at the nucleus, and about two percent higher in F_0 . Isolated, under suitable listening conditions, any of the five differences among the norms for /e/ and / ϵ / are discriminable and

could singly form the basis of a phonological difference. In approximate descending order of audibility, /e/ and /ɛ/ are distinguished by F-pattern, F-pattern movement, duration, F_0 , and level. This order of importance is not fixed. According to Jones (1950, p. 31) some words in Pekinese can be differentiated by F_0 and F-pattern in the vowel, but the phonemic difference must be regarded as primarily one of tone - with a redundant shift in F-pattern. Further (Jones, 1950, p. 192), there are tone languages such as Burmese in which some tones are invariably accompanied by particular voice qualities. For instance, low or falling tone may always be said with a breathy voice, while a vowel with high tone may always have a harsh tone quality. This correspondence (low tone and breathy quality) is not accidental, but neither is it universal.

To summarize, given the assumption of certain possibilities, a first limitation on the Maxi and QNS principles as applied in Chapter 5 is the small set of phonetic attributes that were available. Beside the continuable contoids like trills, laterals, fricatives, and nasal-stops that might form syllabic nuclei and so should be input to Project III, there is a large number of phonetic attributes of vocoids that have not been considered. Just as a two-way distinction of high and low vowels can be made into four vowels by adding a distinction front-back, and into eight vowels with a round-spread distinction, these eight vowels can be made into 32 vowels by independently adding any two laryngeal distinctions, for instance pitch and phonation quality. It is understood that not all these attributes are fully independent, either in production or in perception. Voice quality, F_0 , and voice amplitude seem to be intrinsically related in several ways. Front-back movement of the tongue and lip rounding are interdependent as perceived: i.e., without visual cues they are easily confused. If the limits of independent realization for all the imaginable articulatory or acoustic attributes of vocoids were understood (e.g., $F_2 < 1700$, when $F_1 = 700$; but $F_2 < 2500$, when $F_1 = 250$),

then Project III might have been able to 'predict' that languages ought to be tonal, or that vowel rounding should be made more obvious by having vowels co-vary in duration and rounding so that, *ceteris paribus*, /u/ might be 100 msec long and /i/ 300 msec long.

6.1.2. Caveat: Extra-linguistic influence is not monolithic.

We assume that more than one human tendency may constrain the form of a system. A basic difficulty with QNS and Maxi as applied in Project III, is their isolated operation. Although both QNS and Maxi are reasonable implementations of the common "principle of least effort", as outlined, each operates alone. Surely neither one claims to 'explain' all the sounds and oppositions used in languages, and they could be arranged to operate together. For instance, a routine could run a Maxi decision on the output of QNS or vice versa. Suppose that from among all those articulatory configurations with small gradience, the non-extreme articulatory configurations with favorable combinations of mutual auditory dissimilarity are chosen. Somewhat conversely, consider a rule that chose, from among the articulatory configurations that produce maximally distant sets of auditory events, those that were most stable rather than those closest to neutral.

But 'least effort', as interpreted in QNS and Maxi, is not the only constraining force on vowel systems. Other phonetic or psycholinguistic principles may help explain sound systems.

Coordinate influences.

Three coordinate influences will be reviewed. (1) is "syllable structure", (2) is "consonantal system", and (3) is "information processing".

(1) Syllable Structure. Syllables may be seen in either phonetic or phonological perspective. Stetson (1951) and Ladefoged (1967, Chapter 1) and others have studied the phonetic basis of syllables. Although there is considerable disagreement about the physical cause of syllabic prominence, the phonetic existence of syllabic grouping in

speech production is not often questioned. By one measure or another, segments can be grouped into larger phonetic units. A comprehensive review of the relation of phonetic syllables to phonological syllables may be found in Pike (1967, sections 9.2 and 9.7). Pike argues on linguistic grounds that even though the phonetic syllables may be "reworked" or "overridden" in a phonological system, there are in all languages groups of phonemes that are constrained in their sequential distribution. He argues further that the organization of these distribution potentials can be most clearly understood in relation to the phonetic structure of the syllable, which might predict CV over VC, for instance.

In a related vein, Liberman (1970) suggests that human speech is "special" and that sounds other than speech cannot so efficiently convey information. Liberman focuses on the fact that vocoid portions of a syllable give information about the identity of neighboring consonants and properties of the consonants help to identify the vowel. Liberman argues for the organization of speech encoding from the tacit, metalinguistic principle that segmental information should be rapidly transmitted. Thus, he relates the partial merger of information from adjacent segments into syllabic perceptual units to the limits of temporal resolution in hearing. Liberman says (p. 321) that the complicated encoding of segments into CV syllables "speeds communication by delivering information in parallel".

In sum, there are sequential forces toward certain kinds of structure. Some phonetic features are constrained against prolonged continuation in speech. Some features cannot help being continued over several phonemic segments.

(2) Consonantal System. Aside from sequential structures in the syllable, the possible influence of consonantal phonology on vowel quality systems is obvious. For instance, the number of vowels in a language sometimes seems to be inversely related to the number of consonants. In a 'real' application of Project III (if there could be

such a thing), might not the number of consonants influence the number of vowels to be selected? However, many other aspects of consonant phonology could enter into consideration. The structural relations among consonants may also influence the shape of the vowel system. Jakobson suggests (1968, sec. 25) that the hi-lo dimension is the first one used by children and the one used in unidimensional vowel systems because the hi-lo opposition in vowels is an extension of the consonant-vowel distinction. Since the consonant-vowel distinction must appear before the elaboration of any vowel system, the already operating distinction (C-V; hi-lo; F_1) is used instead of the most basic within-vowel distinction (light-dark; front-back; F_2).

Another way to view the predominance of the hi-lo vowel dimension when only one vowel feature is utilized in the language is to consider the phonological space available for both vowels and consonants as a unitary entity. In the northwest Caucasian languages, cited by Jakobson, the consonantal phonology saturates the phonological space in the front-back and round-spread dimensions. The Caucasian language Ubykh, as analyzed by Paris (1974), has consonants in twenty-three different (primary-second) 'places' of articulation. These 'places' are positions and combinations of positions that differ along the front-back axis in articulation. This leaves only the hi-lo opposition to resist the palatalizing, velarizing, and labializing contextual effects when vowels are put in CVC sequence. The sound system of Iatmul (Staalsen, 1966) provides weak counter evidence to this view, since Iatmul does not saturate the 'light-dark' axis of phonological space, but the unidimensional vowel system of Iatmul is still hi-mid-lo. Like most consonant systems, Iatmul's does include labial, palatal, and velar segments, leaving the whole phonological space below the approximants to the vowels. Staalsen does report (p. 70) extreme allophonic variation along the front-back axis, conditioned by adjacent consonants and naturally leaving only hi-lo unaffected by the sequential consonantal effects.

Stampe (1972) suggests that vowel systems, in isolation should be completely made of non-nasal vocoids because vocoid quality is somewhat 'obscured' by nasalization. If the language under consideration has no nasal consonants, should we assume that nasality could not offer another phonetic parameter to differentiate vowels?

Thus, we conclude with Jakobson (1968, p. 67) that "single sounds must not be treated in an isolated fashion without regard for their place in the sound system."

(3) Information Processing. We may assume that general, non-phonetic, psychological properties of people also could provide inputs to the development of sound systems. These psychological properties (particularly in relation to perception), may be discussed under the heading "information processing". Liberman's explanation for syllabic alternation of consonants and vowels can also be viewed as an appeal to the limits of information processing capacity in the auditory system, and other similar influences can be noted. For instance, Jakobson and Halle (1968) cite the research of Pollack and Ficks (1954) as the prime "weighty argument" in favor of the inherently dichotomous foundation of linguistic sound structure.

First a system of distinctive features based on a mutually implicating relation between the terms of each binary opposition is the optimal code, and it is unwarranted to assume that the speech participants use a more complicated and less economical set of differential criteria in their encoding and decoding operations. Recent experiments have confirmed that multidimensional auditory displays are most easily learned and perceived when 'binary-coded' (Pollack and Ficks, 1954).

The implication is that more dimensions with fewer distinctions per dimension is better than more distinctions along fewer dimensions. That is, eight vowels that differ as dichotomous divisions on three dimensions should be easier to learn and identify than eight points along a single dimension. Several other "limits on our capacity for

processing information" are very clearly presented in Miller's (1956) "Magical number 7 ± 2 ". For instance, there is another significant result reported by Pollack (1953), that has not received much attention from linguists: When identifying sounds that varied only in pitch, the performance of human subjects did not improve appreciably when the range of tones was increased from 100 - 600 Hz to 60 - 14,000 Hz. The efficiency of information transfer in a human system cannot be greatly enhanced by extending the range of a dimension in a display. These psychological principles have far-reaching implications, especially if considered in conjunction with the propositions of section 6.1.1. The F-patterns produced by Lindblom's Maxi procedure are nearly always at the extremes of an F_1 by F_2 plane. Pollack's result suggests that reducing the F_1 by F_2 area used in a vowel system should hardly effect people's ability to learn or operate with the system. Other tendencies of human information processing (e.g., as reported by Eriksen and Hake, 1955) also suggest vowels that are distinguished in several concurrent dimensions would be much easier to identify than extreme vowels with a two-dimensional structure like F_1 by F_2 .

6.1.3 Caveat: Adequate is adequate, but position may be opposition.

This subsection reviews the most basic difficulty with QNS and Maxi; a difficulty with the hypothesis as formulated by Stevens and Lindblom, as well as with our Chapter 5. Even if QNS and Maxi were applied to perfect phonetic models of the complete human vocal capacity, and were integrated with other coordinate influences, they would yield unique, optimal solutions that identify segments as specific points in the phonetic space. There is reason to assume, on the other hand, that adequacy is adequate and that segments are conjunctions of partitions rather than positions.

First, there is little reason to believe that biologically evolved or conventionally developed human systems are or must be optimal. Certainly the range and kinds of variations seen everywhere in nature, and the differences among phonological systems in particular,

necessitate a conclusion that use, development, and feedback do not guarantee perfection. Trubetzkoy (1969, p. 267) characterized particular sound systems as 'economical' or 'wasteful' in terms of functional load. Phonologies do show great diversity, yet they do not seem differentially adaptive. Sound systems may change and develop through the operation of many mechanisms or processes, each of which may be oriented in a determined direction. However, the extant diversity of result suggests that because of multiple influences, the end result is not fixed. Many directions of historical phonological development seem acceptable; none is optimal. Even a universal principle of phonology need not be optimal. In phonology, coding and transmission need only be adequate, but the rigorous formulation of 'adequacy' presents the same difficulties as formulating 'optimality'.

Second, Maxi and QNS produce segment inventories that are precisely positioned in phonetic terms. Chapter 5 optimized vowel targets, it did not identify oppositions or features or dimensions. There is a long tradition in linguistics that views phonological use of phonetic substance as essentially negative. de Saussure (1959, p. 120) states that phonic signifiers "are purely differential and negative ... and in language there are only differences without positive terms". This tradition is continued by Trubetzkoy (1969, p. 72).

It is a basic fact of phonology that the phonemic content of a phoneme depends on the position of this phoneme in the phonemic system and consequently on the structure of that system. Since the systems of distinctive oppositions differ from language to language and from dialect to dialect, the phonemic content of the phonemes also varies according to language and dialect. This difference can also affect the realization of the phonemes.

Trubetzkoy continues after this passage to show that the allophonic possibilities of a phoneme r are conditioned by the whole structural pattern of the language in which an r exists. Likewise, with vowels, what is an u? In a three vowel system /i,a,u/, u is, in this view,

just that vowel that is not i or a, and the allophonic range of variation of /u/ is largely determined by how crowded the vowel space is in a language.

If languages do have "systems of oppositions" rather than "systems of segments", then QNS and Maxi are ill-formulated in Project III. On the other hand, a phonological description in which oppositions are specified without giving positive identification of normal phonetic forms will not allow the comparative description of different languages. It is possible that analysis would reveal that a speaker of English with a heavy foreign accent was operating with the same 'system of oppositions' as used by native speakers. The foreign speaker just did not realize the 'system' with the correct positive terms or the correct allophonics.

We would like to suggest that language-specific phonologies must be developed with both positive and negative terms. Phonemes that were identified solely in positive phonetic terms would give no clue as to their possibilities in allophonics and in variation. A partitioning of the English phonological space that was sufficient to isolate /o/, for instance, would probably leave a slot that could be realized by some very non-English sounds.

To Summarize: Sections 6.1.1,2,3 indicate that our Project III may not work in the large as it does in the small.

6.2 Conclusions

The three most important findings from the research reported in this dissertation are the following:

(1) Those aspects of the auditory quality of voiced vocoids that are controlled by the formant frequencies cannot be modeled as an additive combination of scaled formant frequencies.

(2) On the contrary, some representation of formant intensity levels must be represented in order to model auditory dissimilarity among voiced vocoids. It is suggested that a polynomial combination

rule on the formant frequencies and formant levels will allow satisfactory modeling of our similarity data.

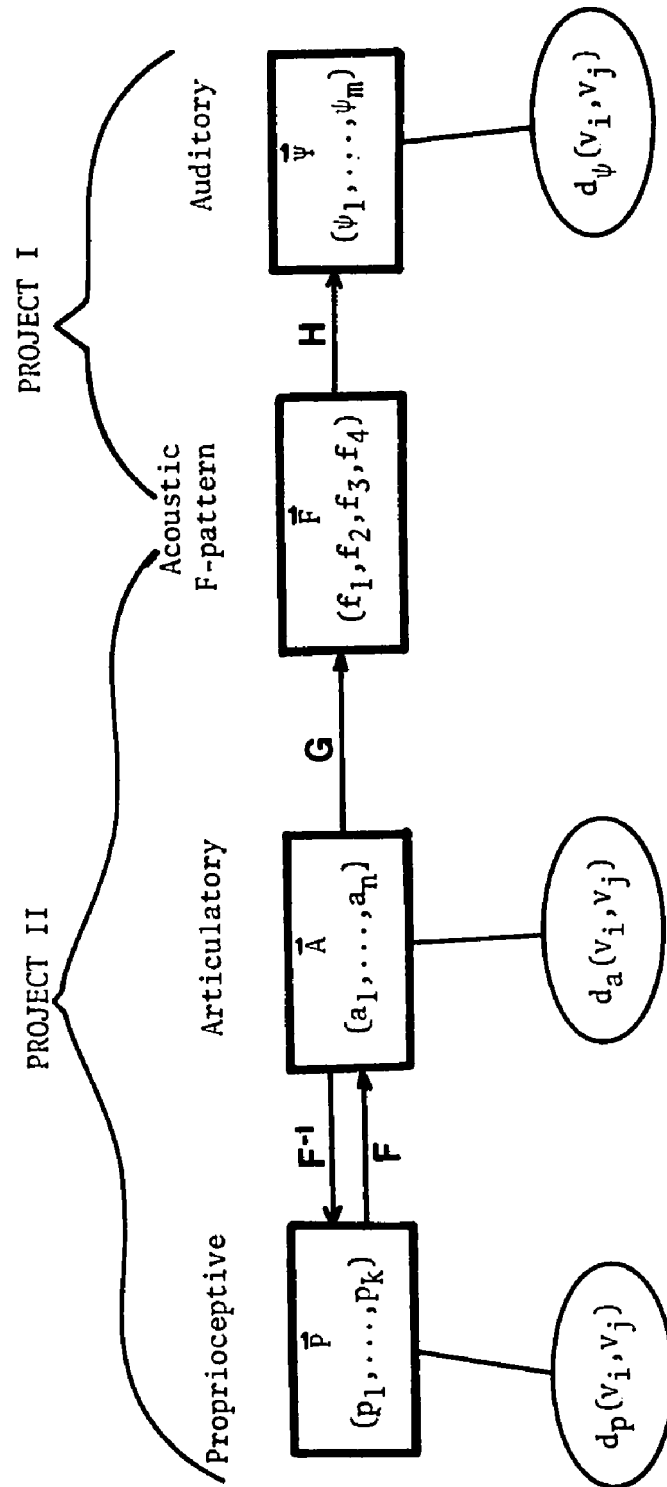
(3) There is sufficient uncertainty in the formulation and verification of the "Quantal Nature of Speech" and Lindblom's "Maximization" hypothesis to question whether either may form an important explanatory principle in phonology.

Findings (1) and (2) are developed in Chapter 3 and (3) is developed in section 1.4, Chapter 5 and section 6.1 above.

Project III was the application of the Quantal Nature of Speech and the Maximization hypotheses to our model of the transmission of vocoids in the phonetic medium. Since the parts of that model vary considerably in their level of verification, it is of some interest to consider whether or not one part of the model of the phonetic channel can have overriding influence on the outcome of Project III. Although this would be at odds conceptually with the discussion in sections 1.3 and 1.4, the predominance of one phonetic stage could be a function of the state of sophistication in the discipline.

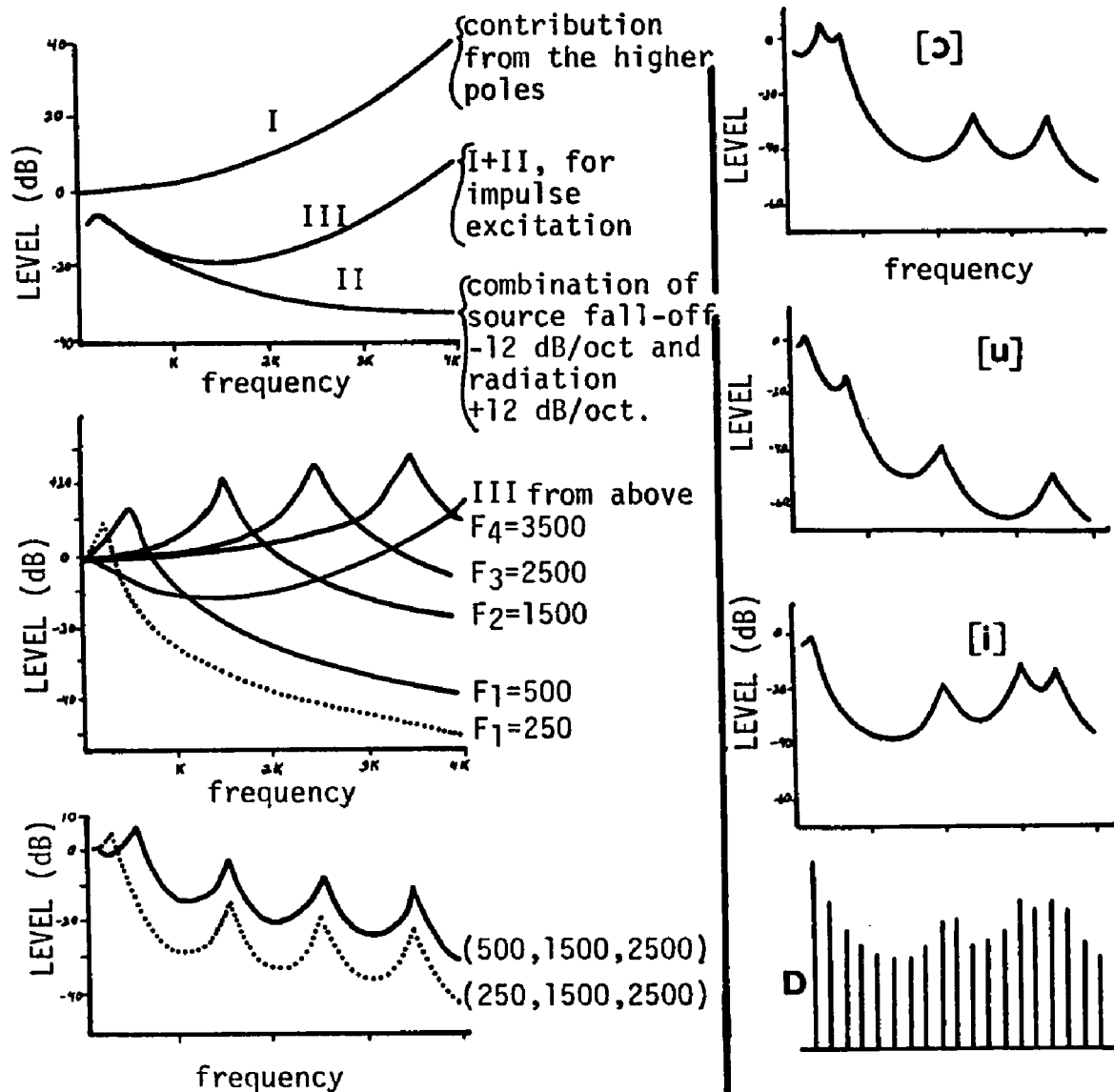
The parametric structure of the vocal tract model in Chapter 4 is extremely limited. Furthermore, the operation of the vocal tract model not only determines the relation of input to output in the channel, but it determines the exact size and shape of the output range. The results of QNS did not change appreciably when the auditory distance and the articulatory distances were altered. Given a "95% of maximal" procedure on a space of possible vocoids, the distance formula in Maxi will have little effect. The shape of the space can change the outcome very markedly.

If one wanted to pursue Project III further, the most basic element that needs improvement is the model of articulation. Of secondary importance would be the testing and refinement of our auditory distance using finer psychoacoustic methods. However, as the Caveats section (6.1) suggests, Project III deserves radical reformulation or abandonment.



Functional representation of phonetic channel, $H(G(F(\vec{P}))) = \vec{\Psi}$, for vocoids.
 Distance functions on pairs of vectors are shown at three stages.

FIGURE 1.1-1



Factors contributing to the relative level of F-patterns

D is the approximate spectrum envelope of the transfer function [i] above as excited by an impulse train at 200 Hz.

FIGURE 1.2-1: THE DETERMINATION OF SPECTRUM LEVELS IN RELATION TO F-PATTERNS
(adapted from Fant, 1970.)

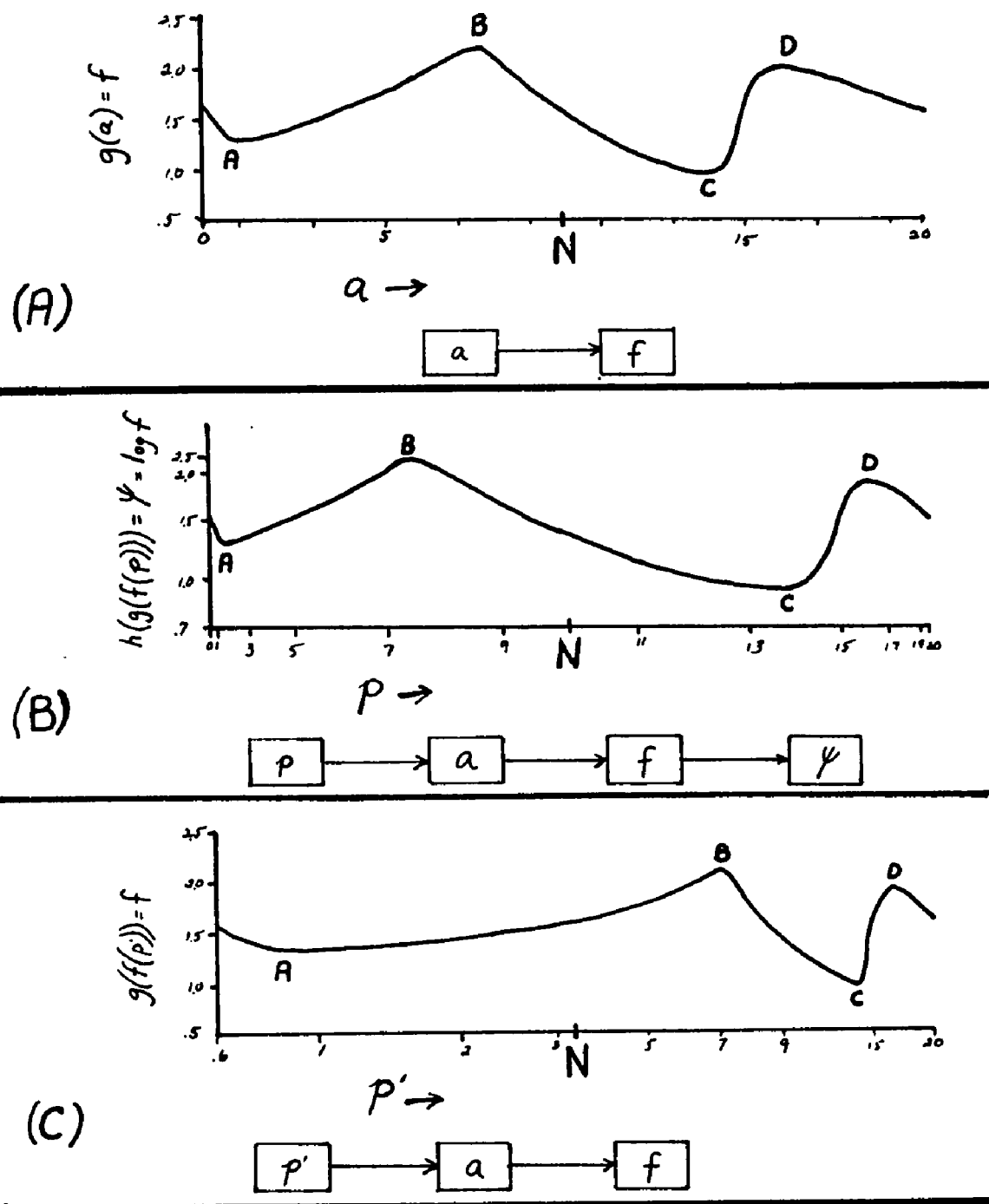


FIGURE 1.4-1 THREE SCALINGS OF THE FUNCTION $g(a) = f$, WHERE a IS THE PLACE OF CONSTRICTION IN cm. ALONG A SIMPLE VOCAL TRACT MODEL AND f IS THE FREQUENCY OF THE SECOND FORMANT IN kHz.
Adapted from Fant, 1970 p.76.

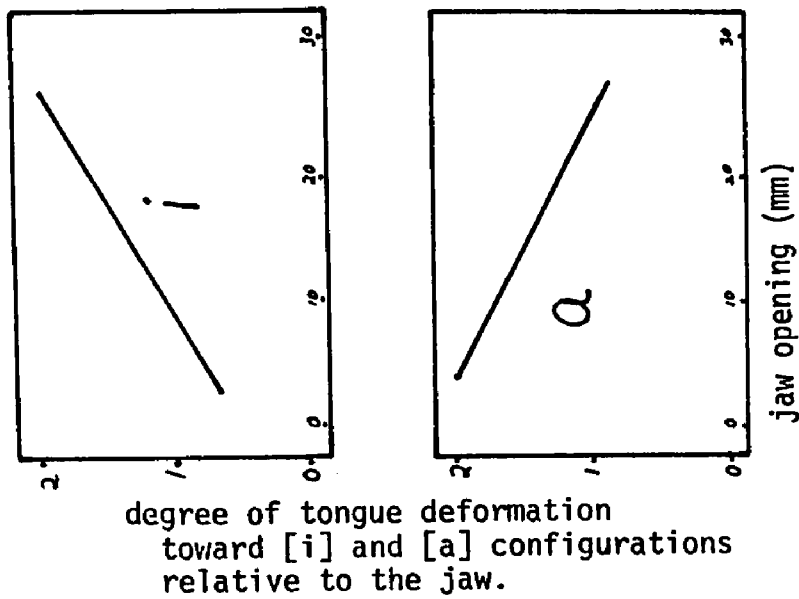


FIGURE 1.4-2: SETS OF (JAW OPENING, TONGUE DEFORMATION) VALUES THAT PRODUCE ACCEPTABLE [i] and [a] in the LINDBLOM-SUNDBERG VOCAL TRACT MODEL. See Lindblom and Sundberg, 1971.

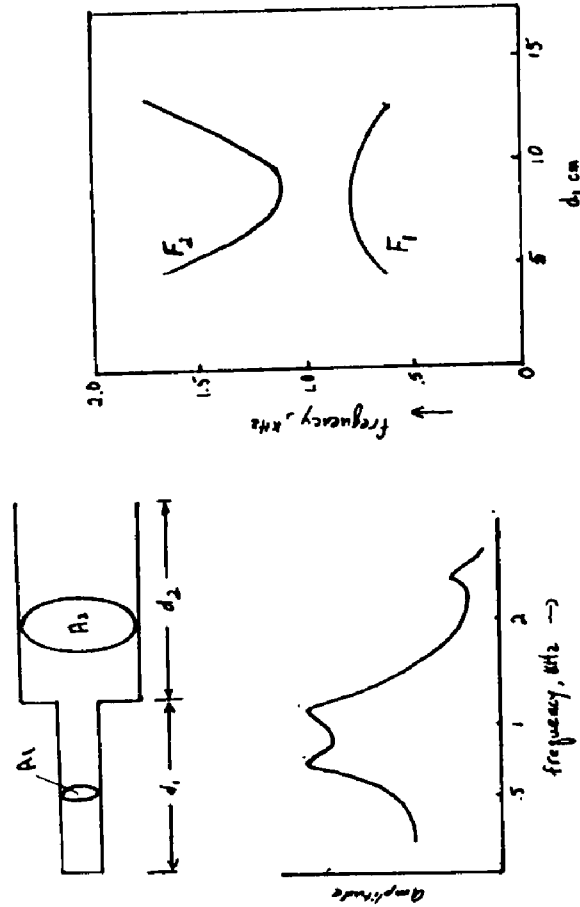
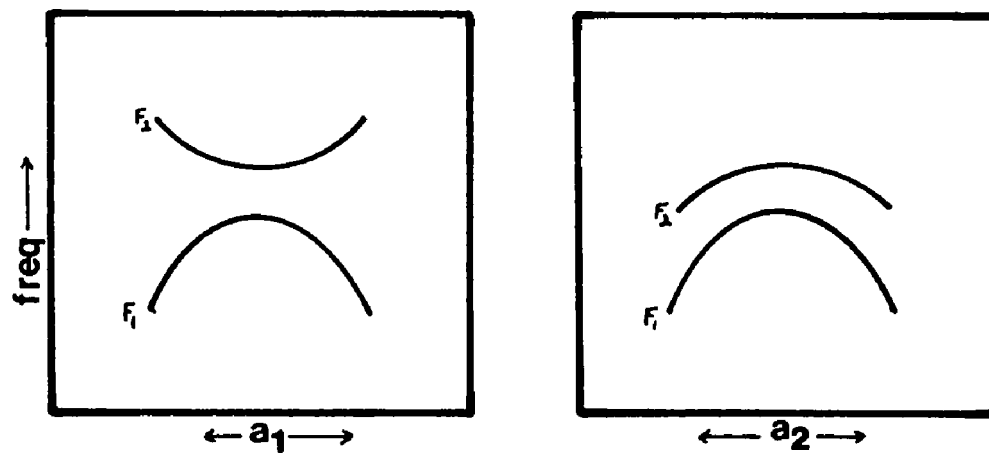


FIGURE 1.4-3: (a) TWO TUBE APPROXIMATION OF VOCAL TRACT CONFIGURATION FOR THE VOWEL /a/, and THE APPROXIMATE FORM OF THE GENERATED SPECTRUM ENVELOPE.

(b) EXAMPLE OF RELATION BETWEEN F_1, F_2 AND THE LENGTH d_1 OF THE BACK TUBE OF THE TWO TUBE RESONATOR IN (a). THE AREAS A_1 AND A_2 ARE 0.2 AND 3.0 cm^2 , RESPECTIVELY. From Stevens, 1972.

FIGURE 1.4-4: ILLUSTRATING HOW L_2 CAN CHANGE THE ORDER OF GRADIENTS.



$$g(a_1, a_2) = (f_1, f_2)$$

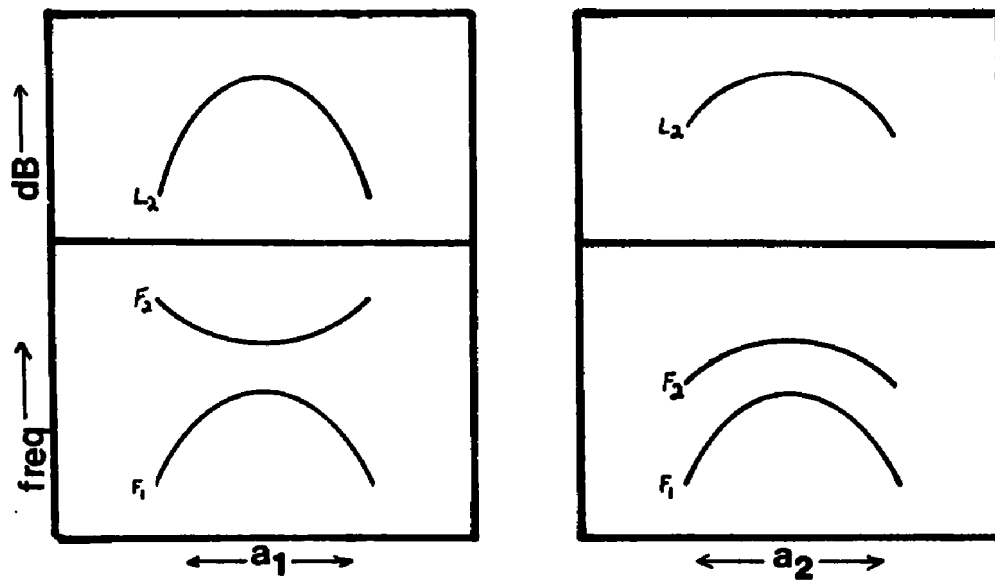
$$\nabla g(a_1) \cong \nabla g(a_2)$$

1.4-4a

1.4-4b

$$g(a_1, a_2) = (f_1, f_2, L_2)$$

$$\nabla g(a_1) \neq \nabla g(a_2)$$



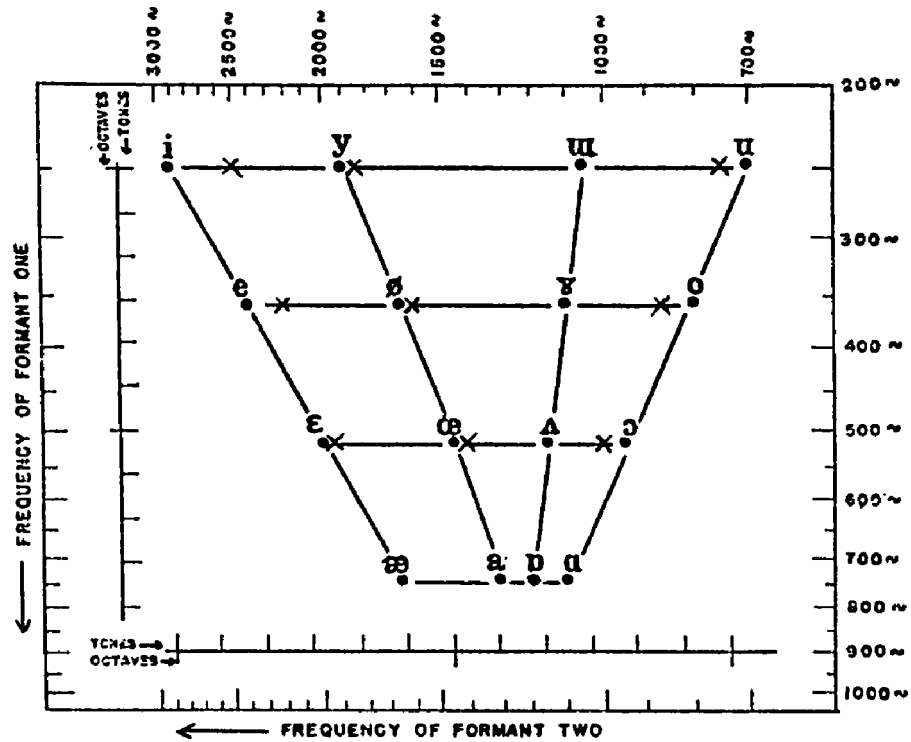


FIGURE 2.2-1: THE BEST TWO-FORMANT MATCHES TO THE 16 CARDINAL VOWELS. THE X'S SHOW MEASURED F_1, F_2 FOR FRENCH VOWELS.

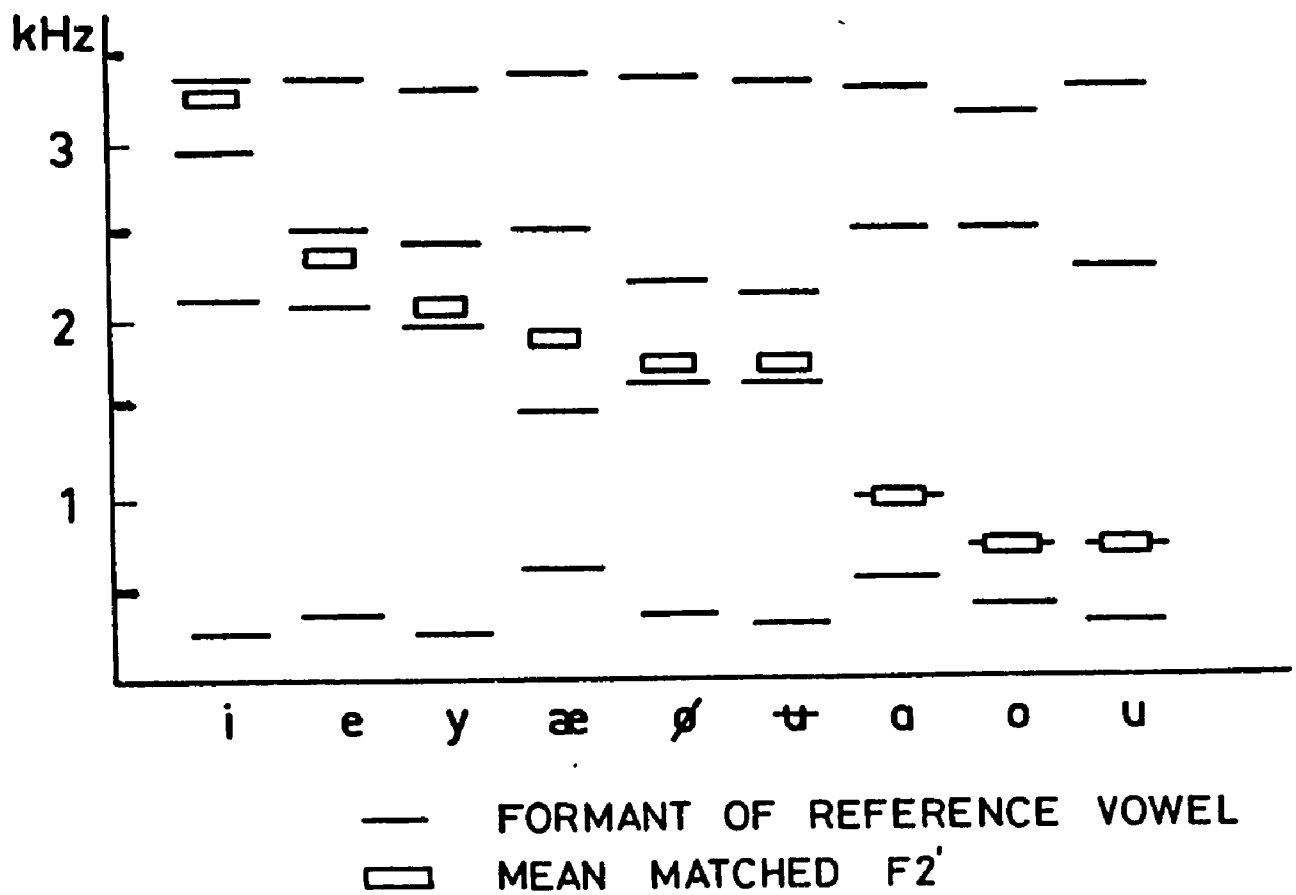
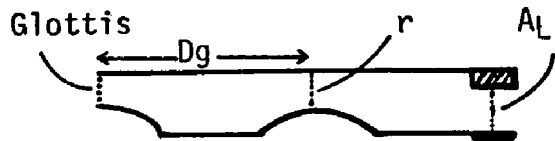


FIGURE 2.2-2: EMPIRICALLY DETERMINED F2' FOR SWEDISH VOWELS.
 From Carlson, Granström, and Fant, 1970.



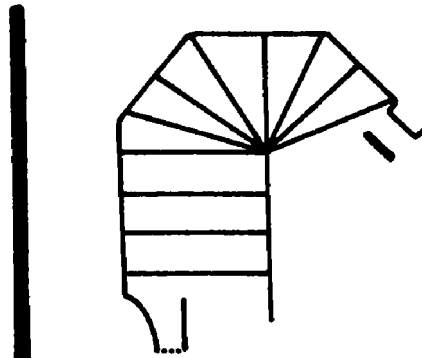
3-Parameter Vocal Tract Model

D_g : distance from glottis

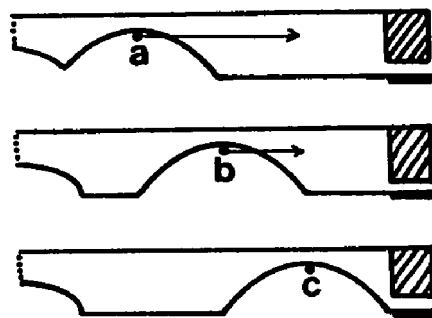
r : radius at constriction

A_L : area at lips

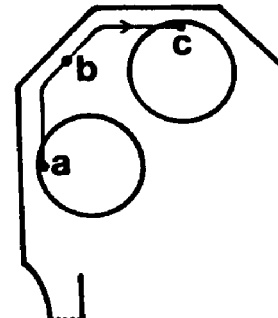
Initial (a), final (c), and a medial (b) position in a simple parametric route from [ɔ] to [y]. A_L and r are fixed, just D_g changes. (c) $\approx$ [u].



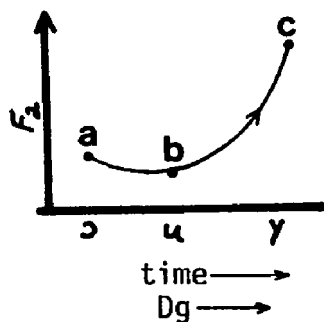
Schematic vocal tract in mid-sagittal section with semi-polar coordinates



becomes



acoustic result



CORRECT ARTICULATORY AND ACOUSTIC PATH FROM [ɔ] TO [y].

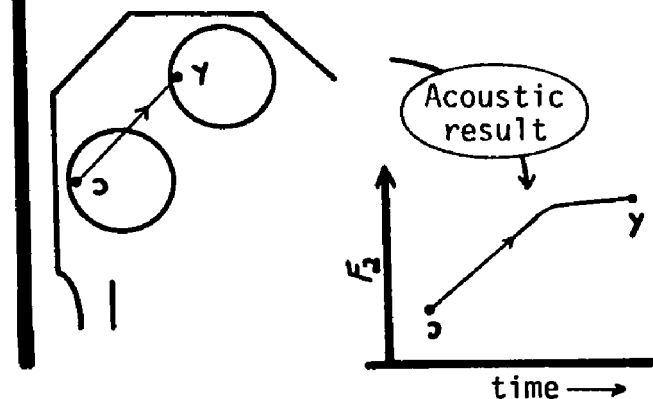


FIGURE 2.4-1 [ɔ] - [y] PATHWAY IN A 3-PARAMETER VOCAL TRACT MODEL

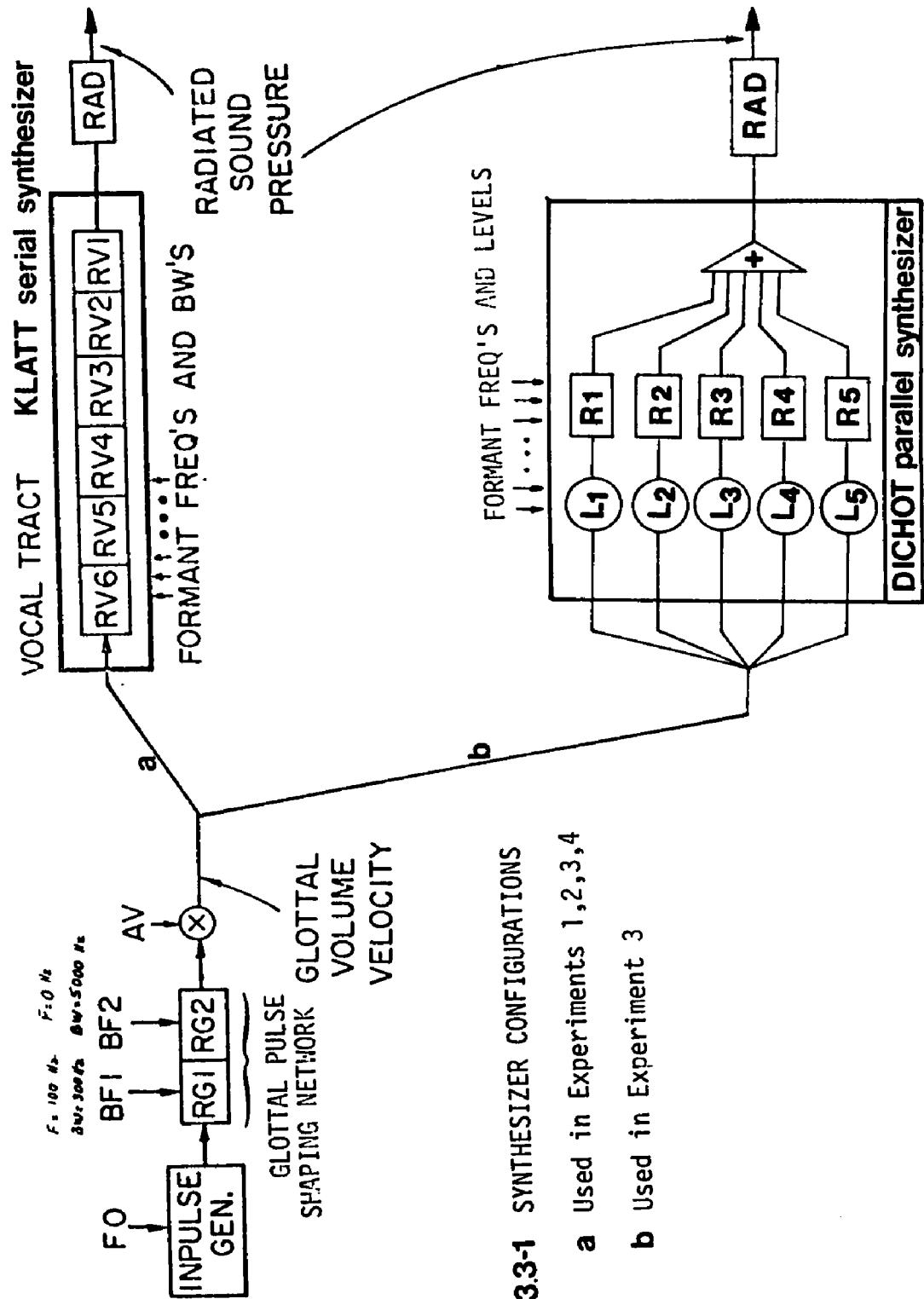
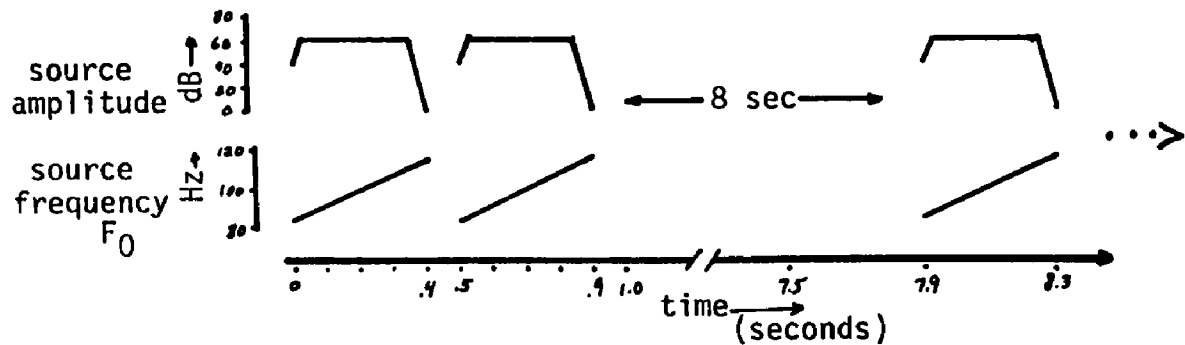


FIGURE 3.3-1 SYNTHESIZER CONFIGURATIONS

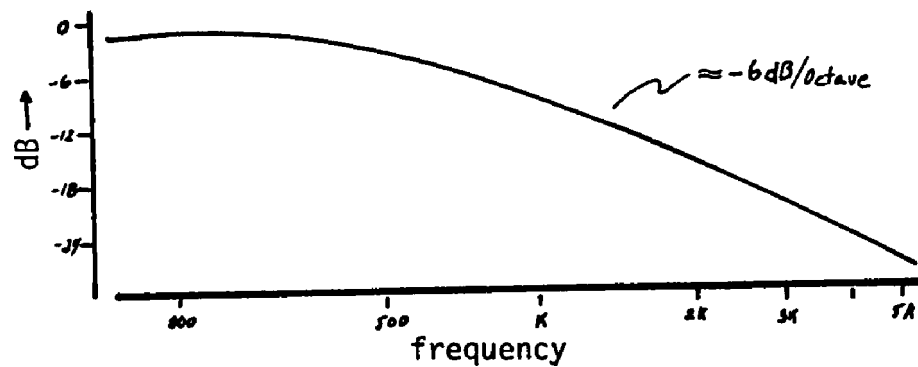
a Used in Experiments 1,2,3,4

b Used in Experiment 3

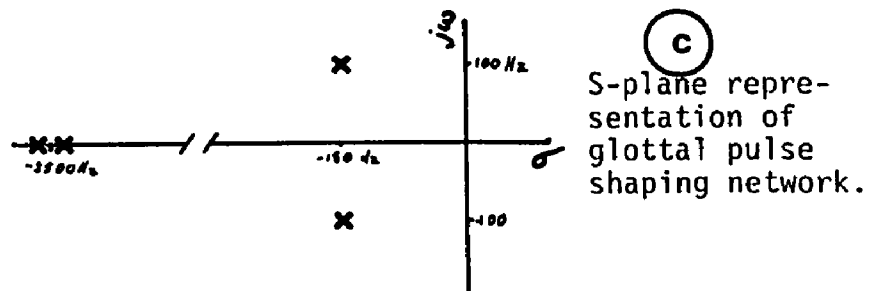
FIGURE 3.3-2 UNIFORM PROPERTIES OF STIMULI IN EXPERIMENTS 2,3,4.

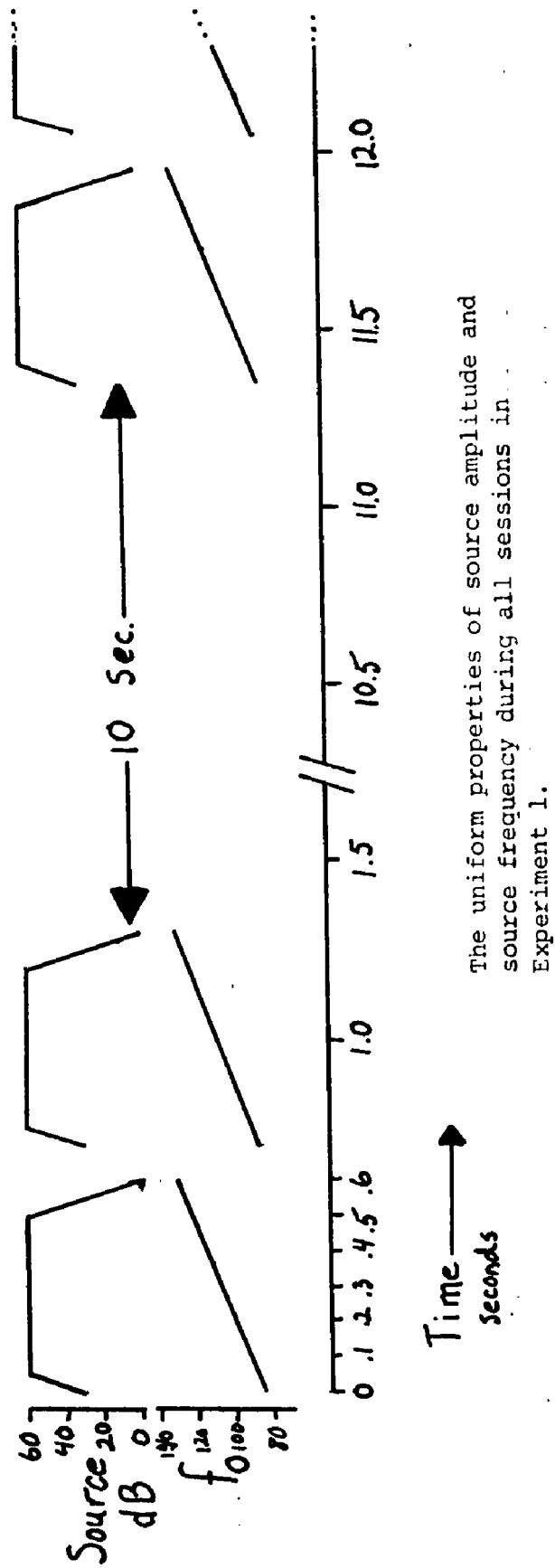


- (a)** Usual temporal properties of source amplitude and frequency for stimulus pairs



- (b)** Combined effect of glottal pulse shaping network and RAD approximates -6dB/octave.





The uniform properties of source amplitude and source frequency during all sessions in Experiment 1.

Figure 3.4-1

FIGURE 3.4.2

A three dimensional representation of the stimuli of Experiment 1.
 The dimensions represent formant frequencies, logarithmically scaled.
 The dark points are the hub stimuli, 555, 112, 221, 332.

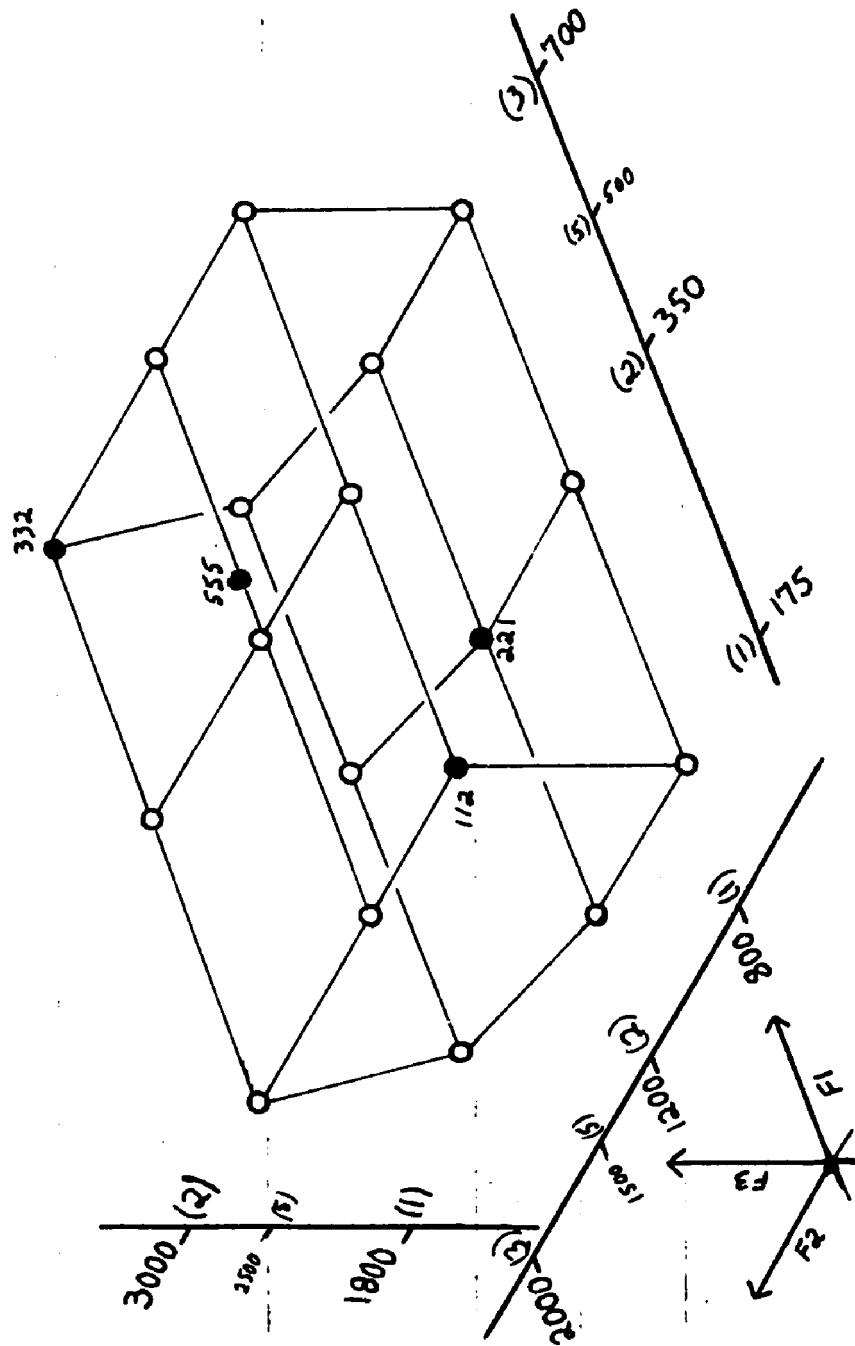
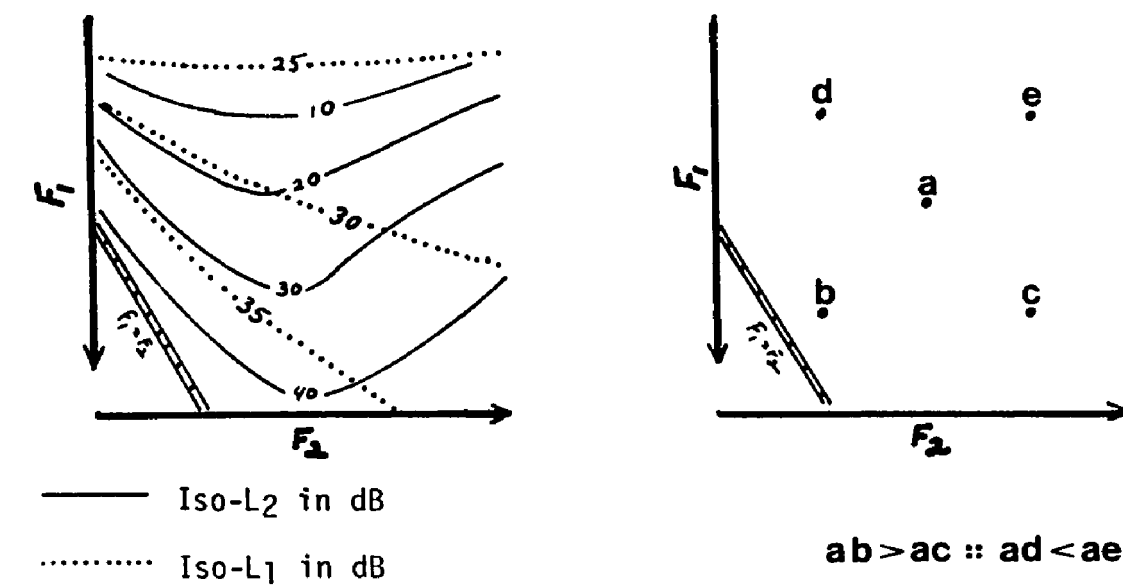


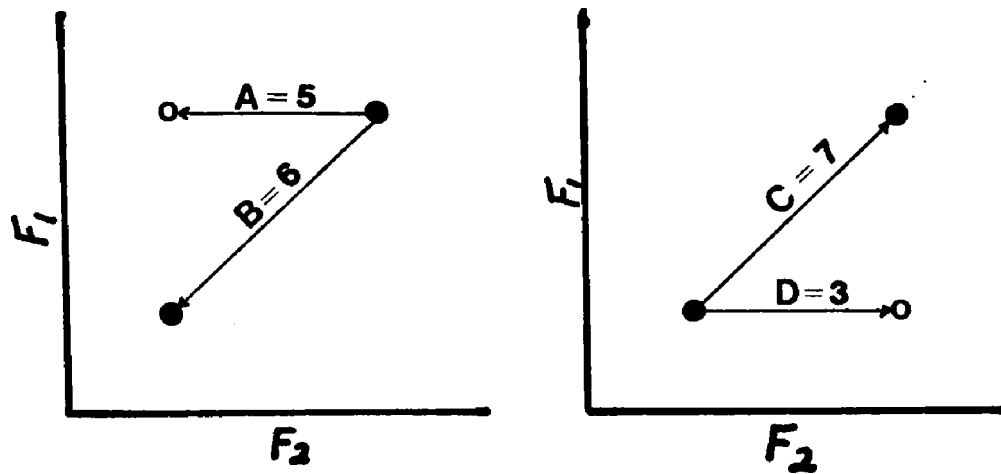
FIGURE 3.4-3: AN EXAMPLE OF AN EXPECTED VIOLATION OF DECOMPOSABILITY (3a) CAUSED BY PATTERNS OF L_1 AND L_2 DIFFERENCES.



	a	b	a	c	a	d	a	e
L ₁	30	39	30	32	30	28	30	26
L ₂	22	36	22	36	22	11	22	16

	a	b	a	c	a	d	a	e
L₁	30	39	30	32	30	28	30	26
L₂	22	36	22	36	22	11	22	16

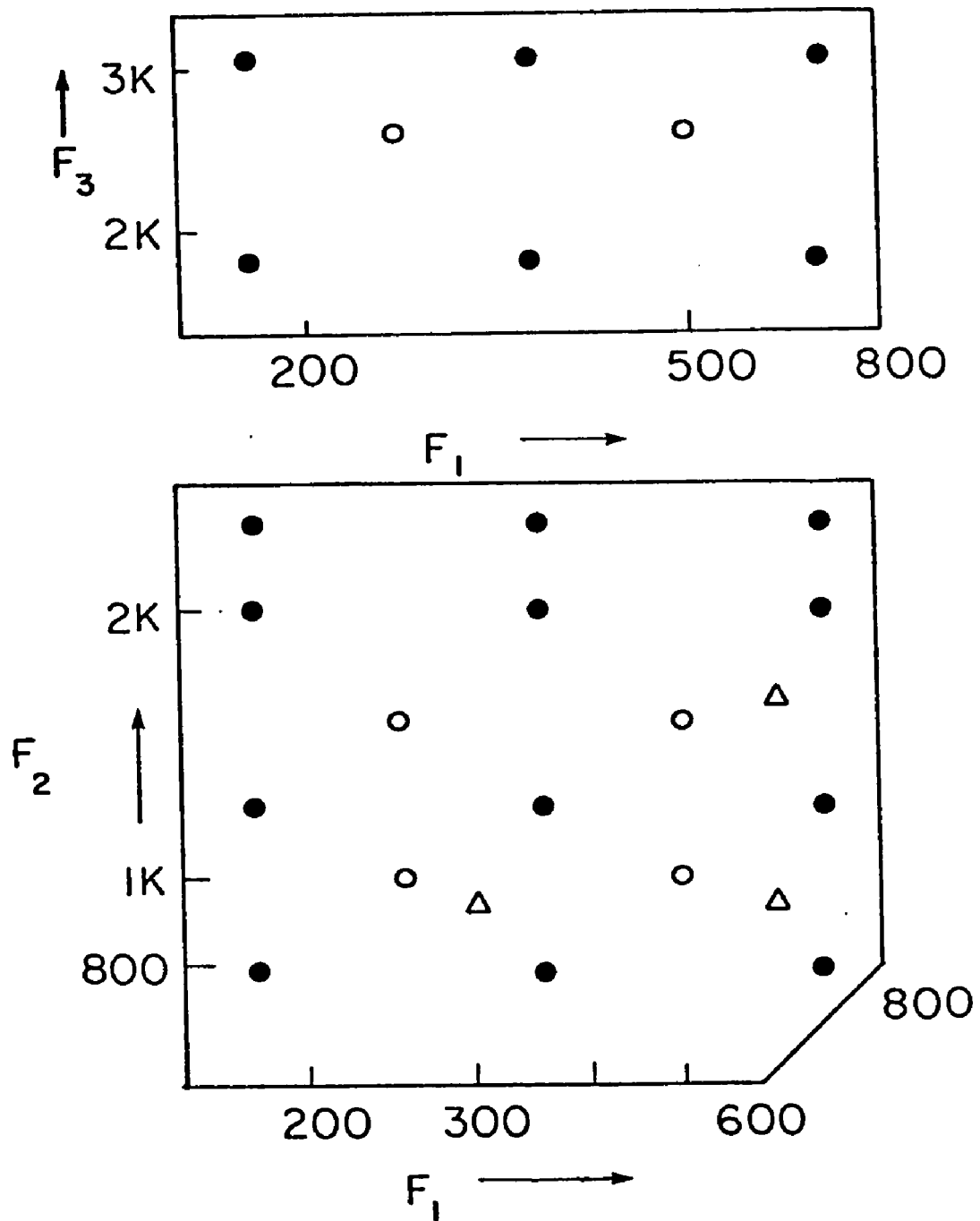
FIGURE 3.4-4 INDICATIVE VIOLATION OF (E1)
FROM DECOMPOSABILITY



$$C > B \quad :: \quad A > D$$

Example from Experiment 1:

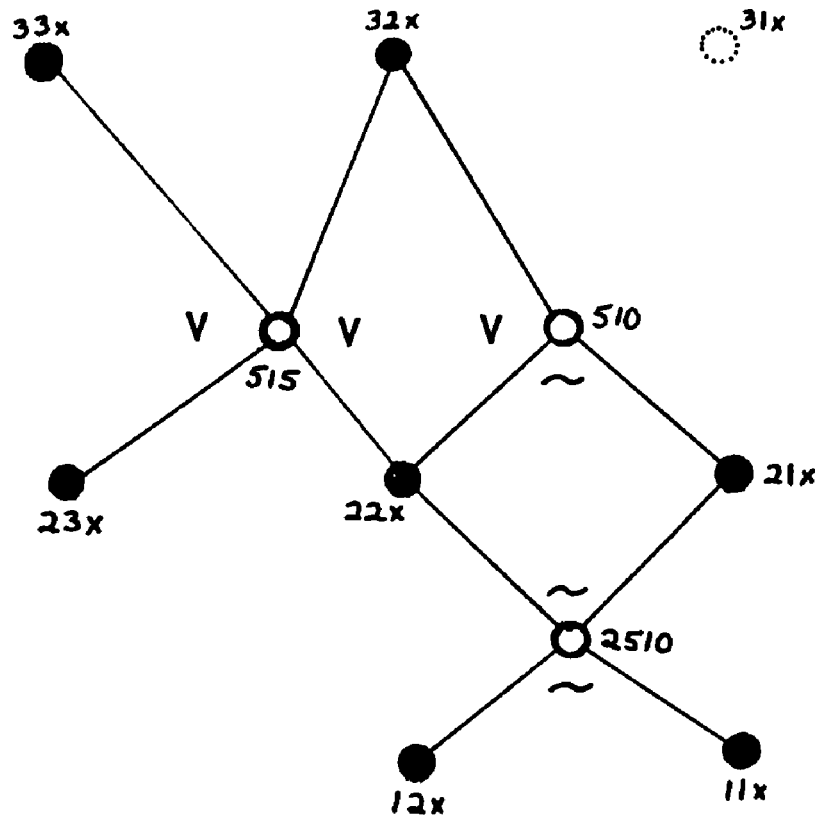
$$112,221 > 221,112 \quad :: \quad 221,211 < 112,122$$



- 21 POINTS IN $F_1 \times F_2$
- 4 HUB POINTS IN $F_1 \times F_2$ USED IN EXPERIMENT 2a
- △ 3 HUB POINTS IN $F_1 \times F_2$ used in Experiment 2b

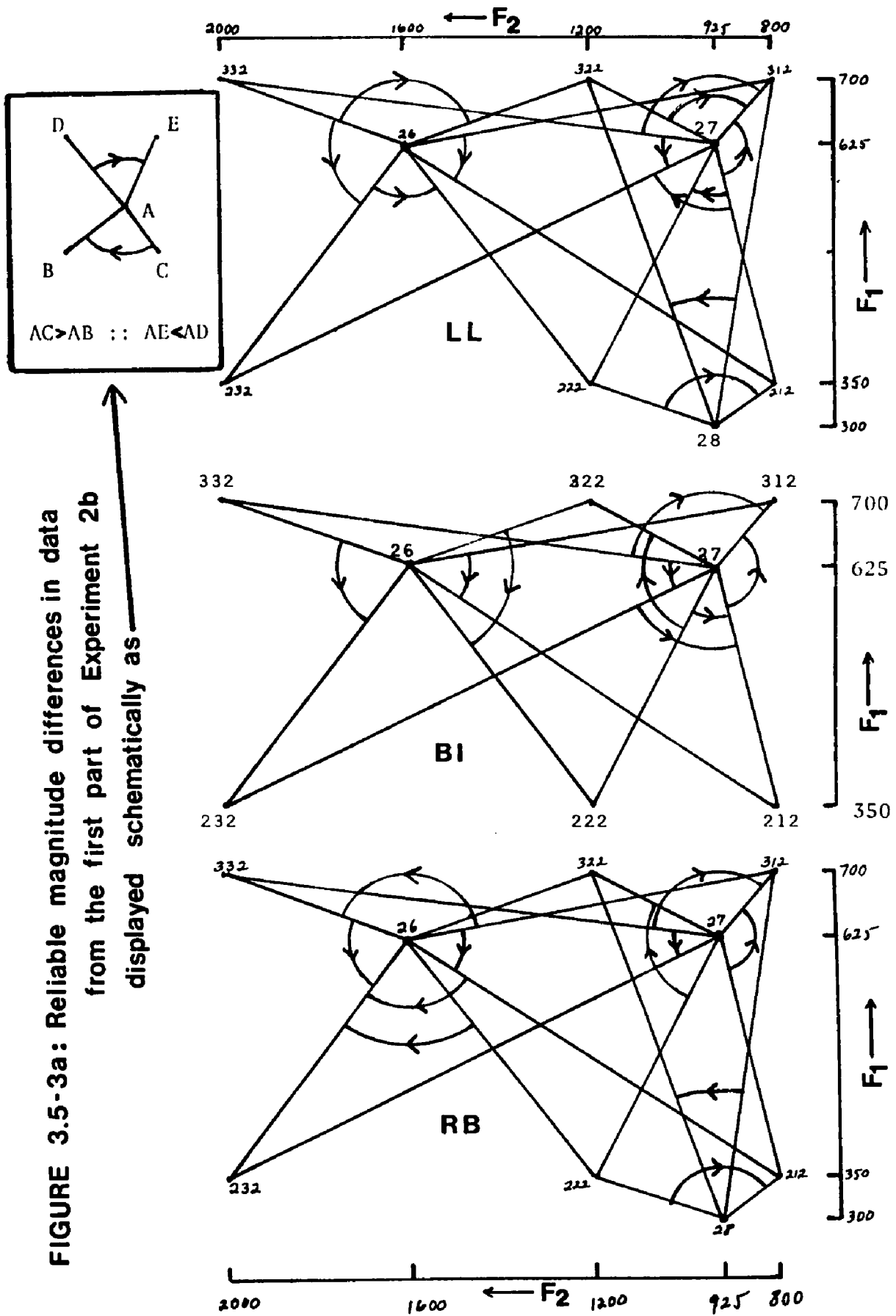
FIGURE 3.5-1

FIGURE 3.5-2 SOME STIMULI OF EXPERIMENT 2a
DISPLAYED IN APPROXIMATE
AUDITORY SCALE



Part of the stimulus space of Experiment 2a seen in an F_1 by F_2 projection with the axes scaled approximately to fit the listeners' data for all stimuli except 31x.

Notice that under an ORDINAL test, such as (E3), no matter how large the distance from a hub to 31x, no violation can be formed.



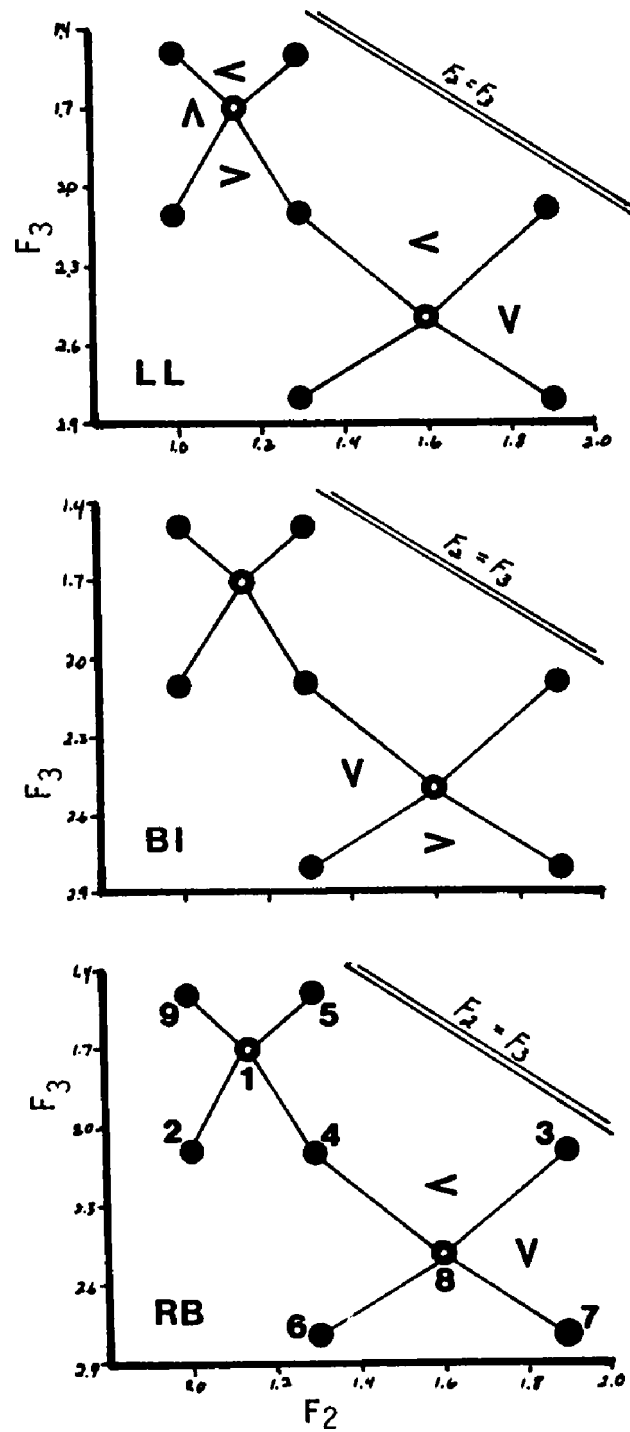
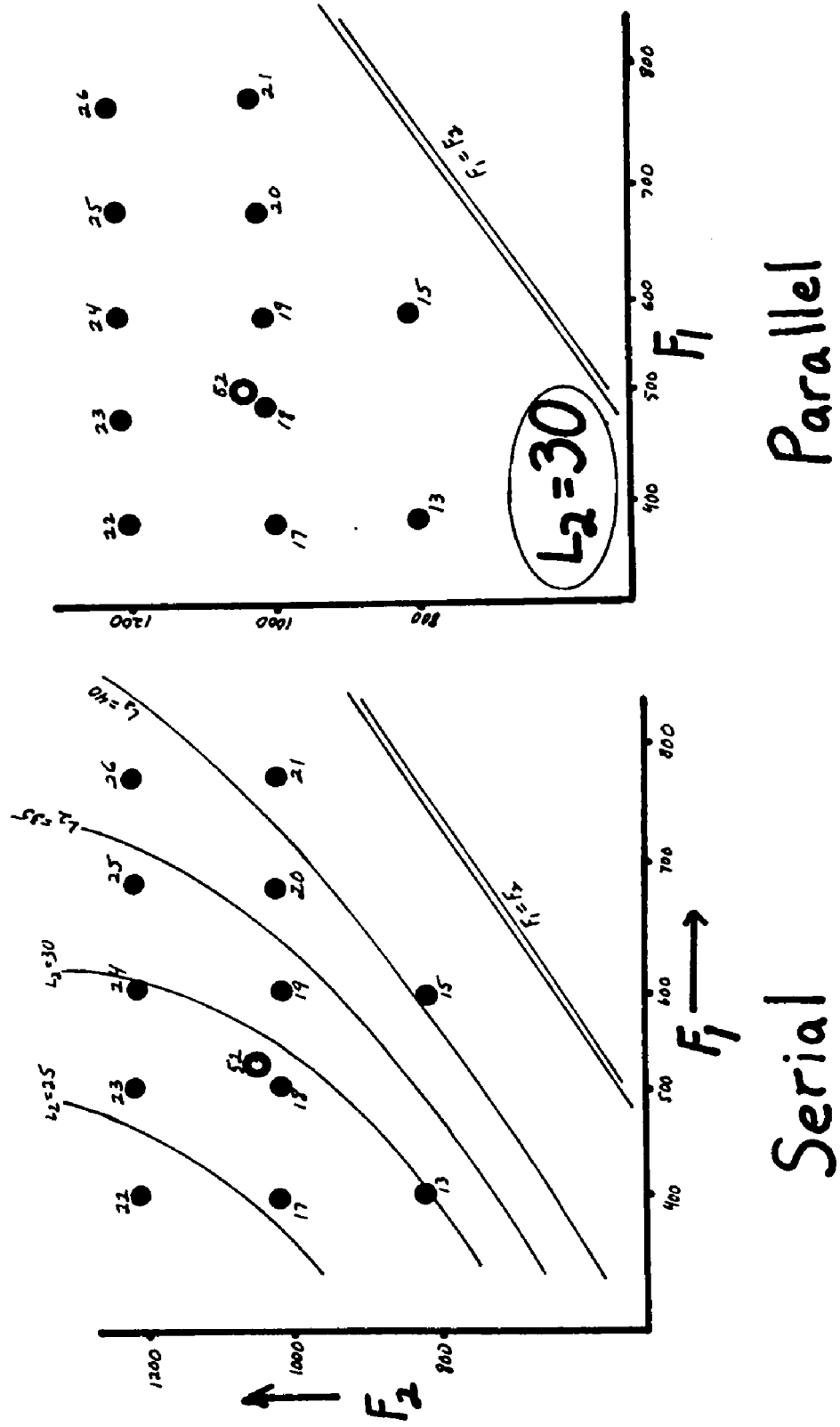


FIGURE 3.5-3b STRONG ORDERS IN THE DATA FROM THE SECOND PART OF EXPERIMENT 2b.

FIGURE 3.6-1: Stimuli for Experiment 3



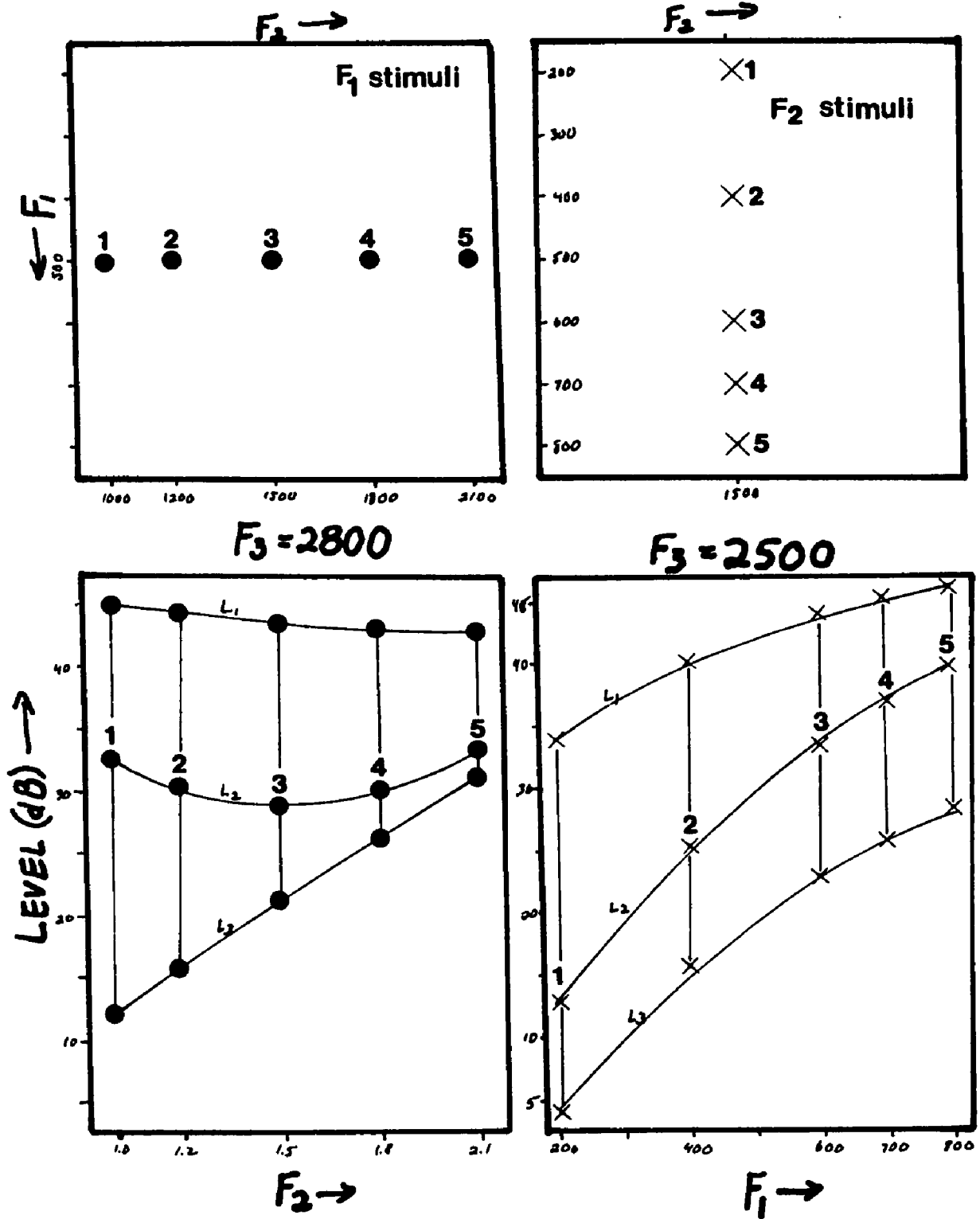


FIGURE 3.7-1 STIMULUS PROPERTIES IN EXPERIMENT 4

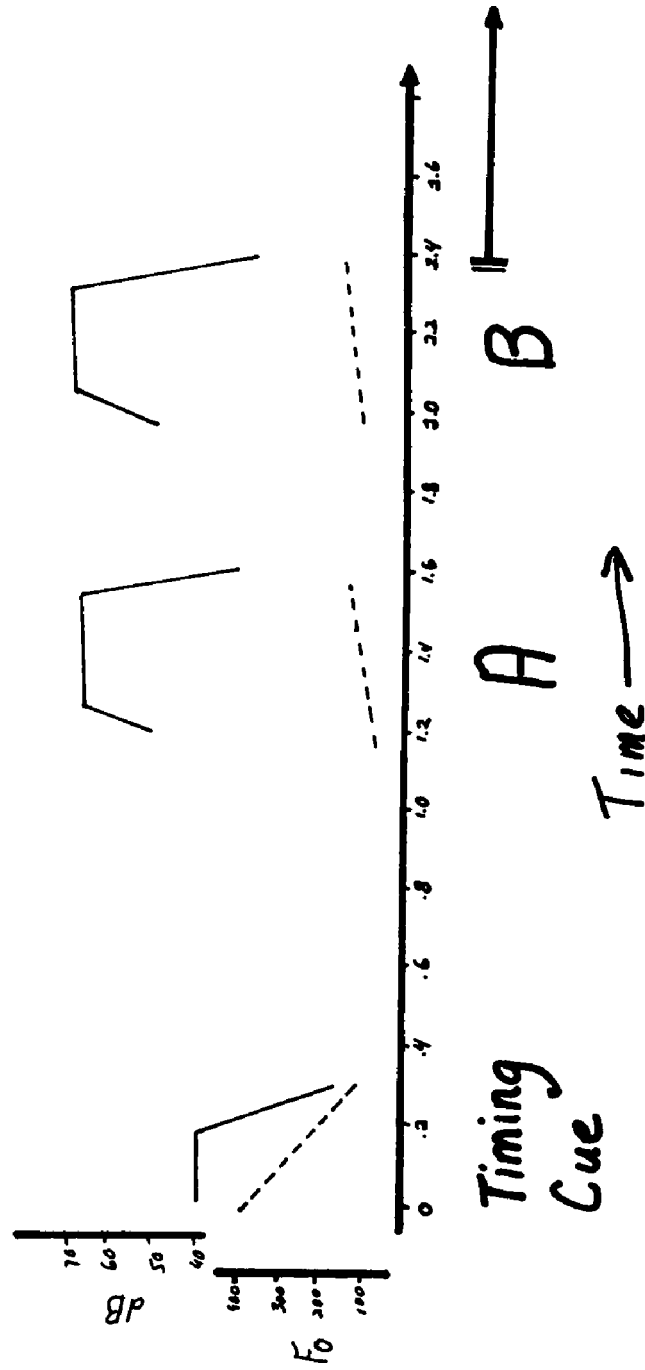
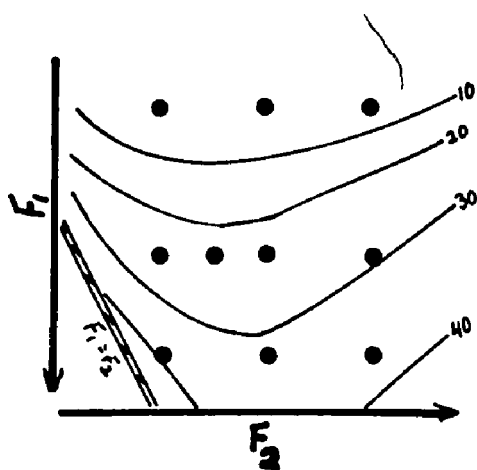
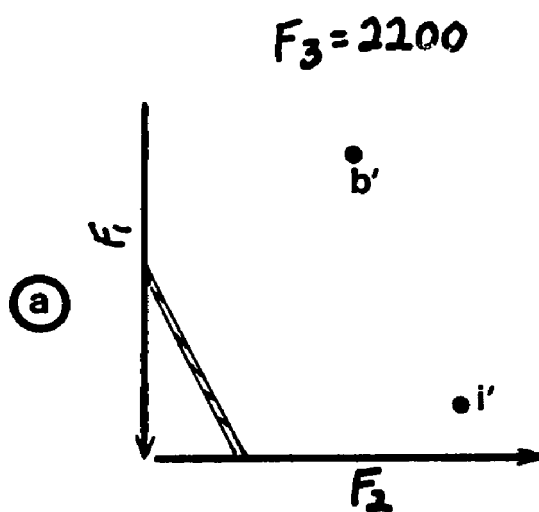
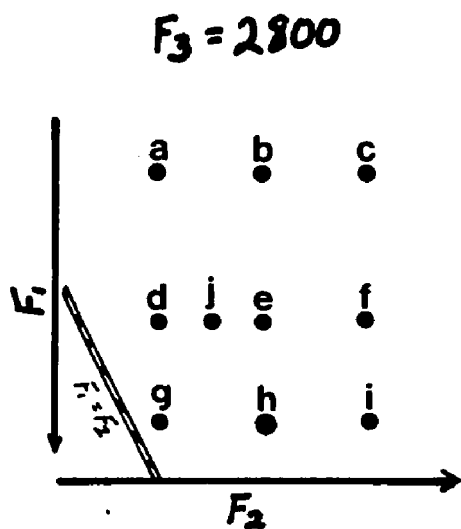
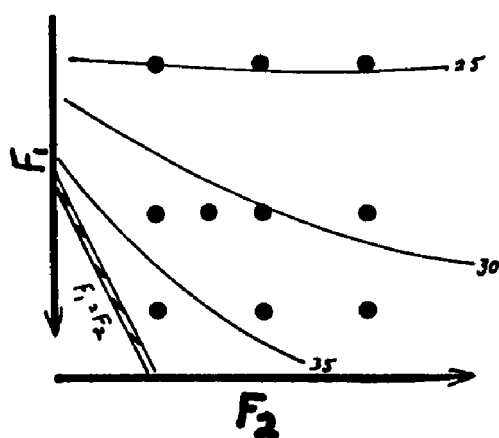


FIGURE 3.7-2: CONSTANT SOURCE PROPERTIES AND TEMPORAL RELATIONS IN EXPERIMENT 4.



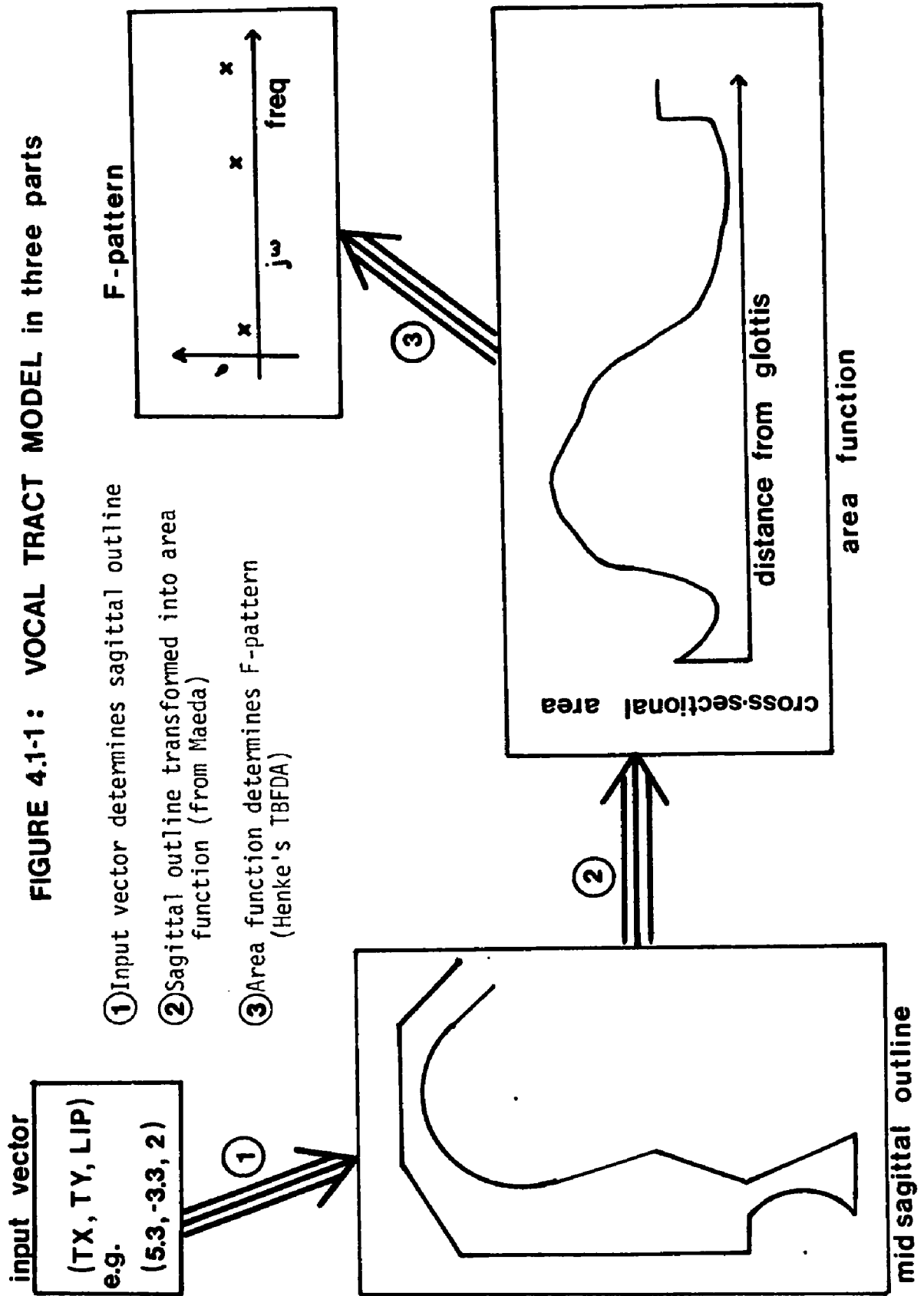
(b) Iso- L_2 every 10 dB



(c) Iso- L_1 every 5 dB

FIGURE 3.8-1_{a,b,c}

FIGURE 4.1-1: VOCAL TRACT MODEL in three parts



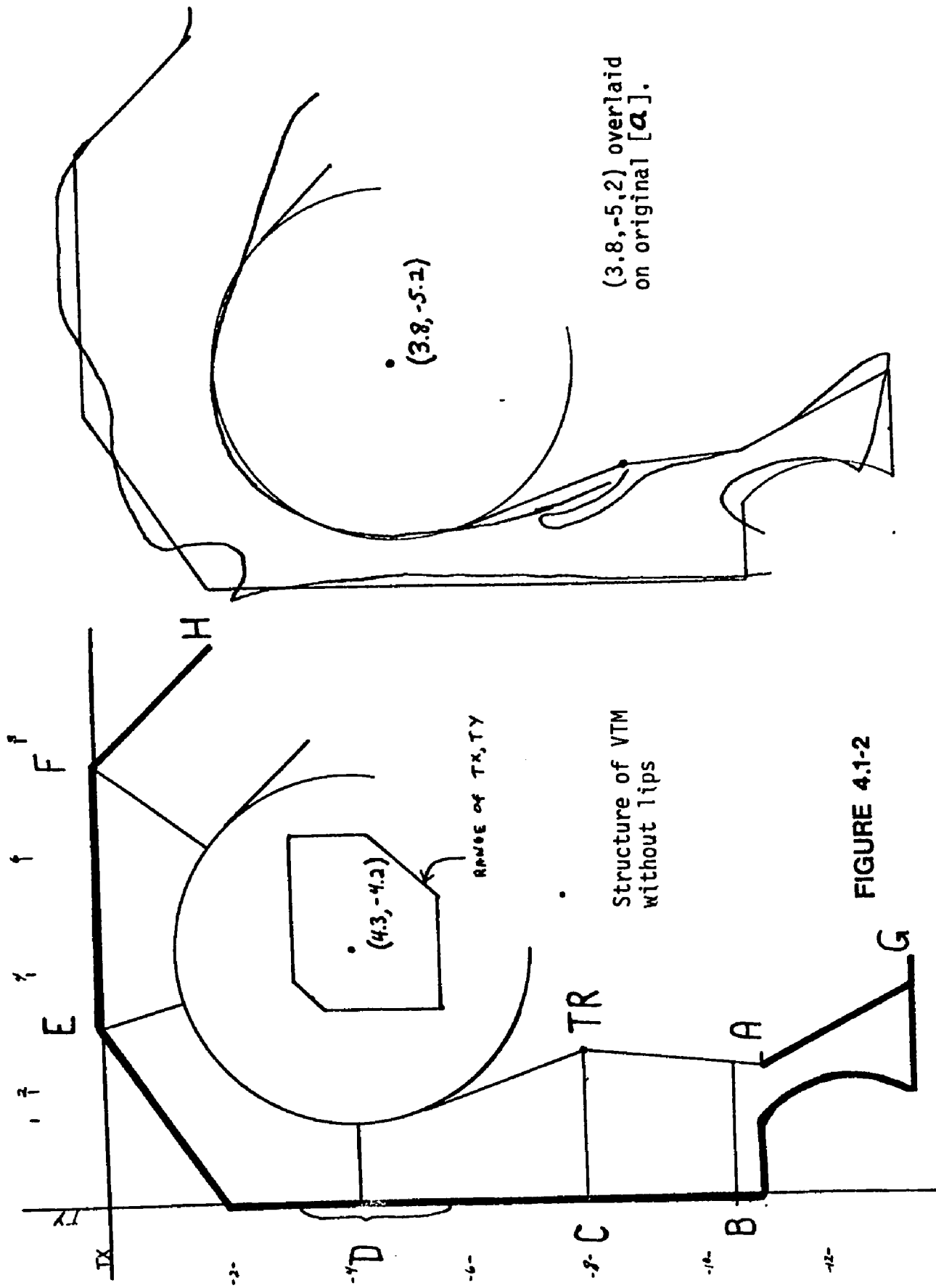
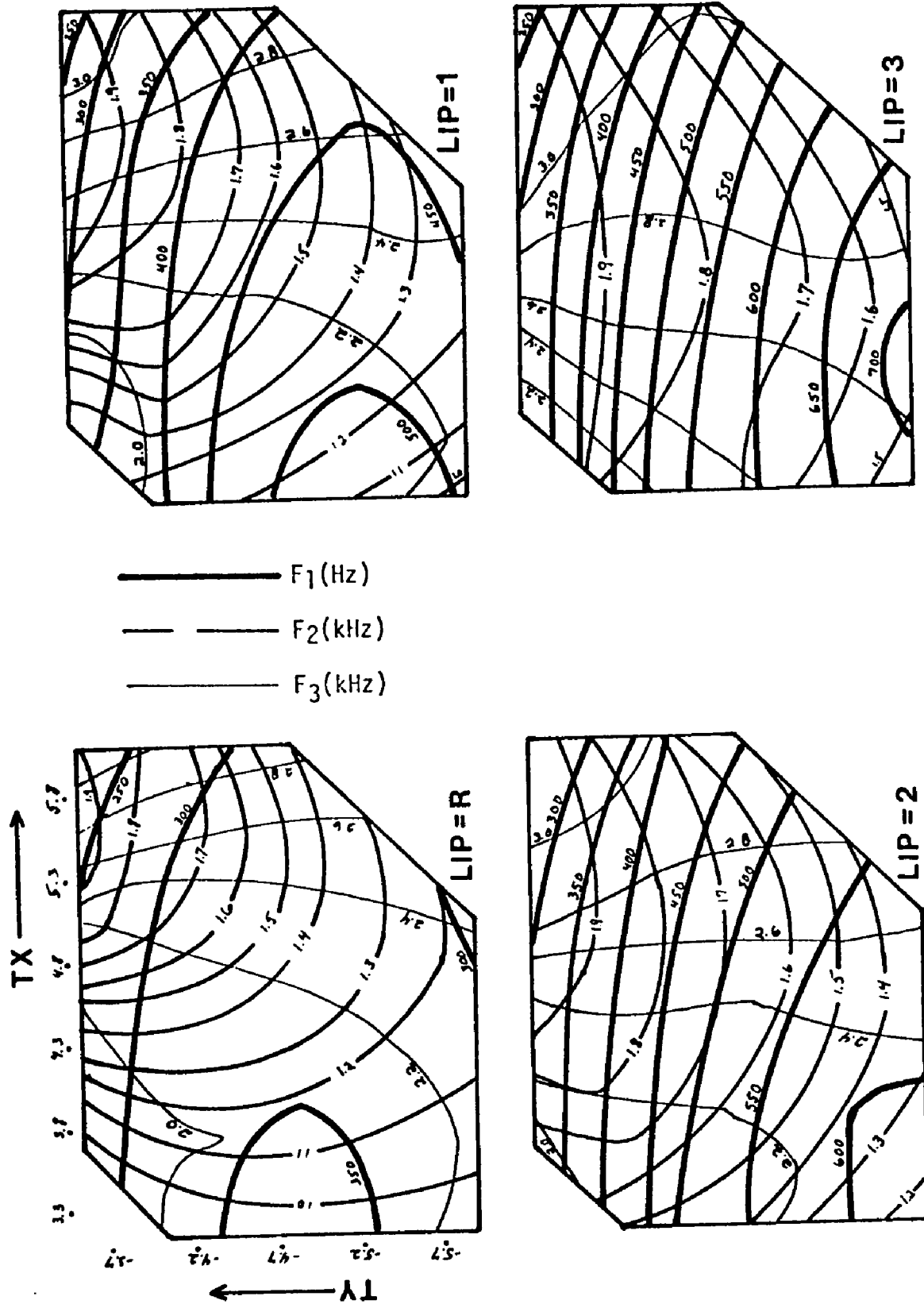


FIGURE 4.1-2

FIGURE 4.1-3: FORMANT FREQUENCIES AS A FUNCTION OF TX AND TY
AT LIP = R,1,2,3 IN OUR VOCAL TRACT MODEL



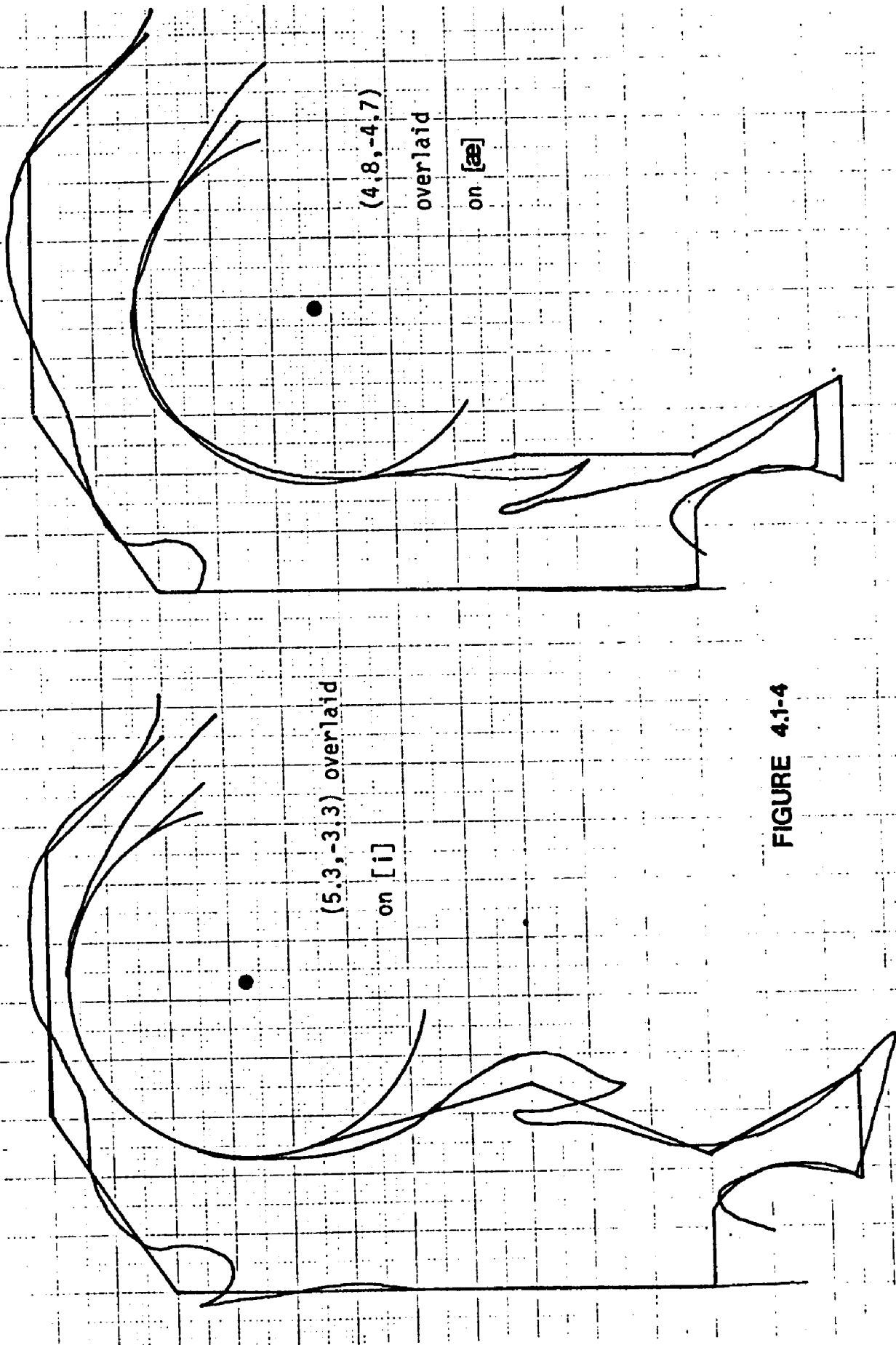


FIGURE 4.1-4

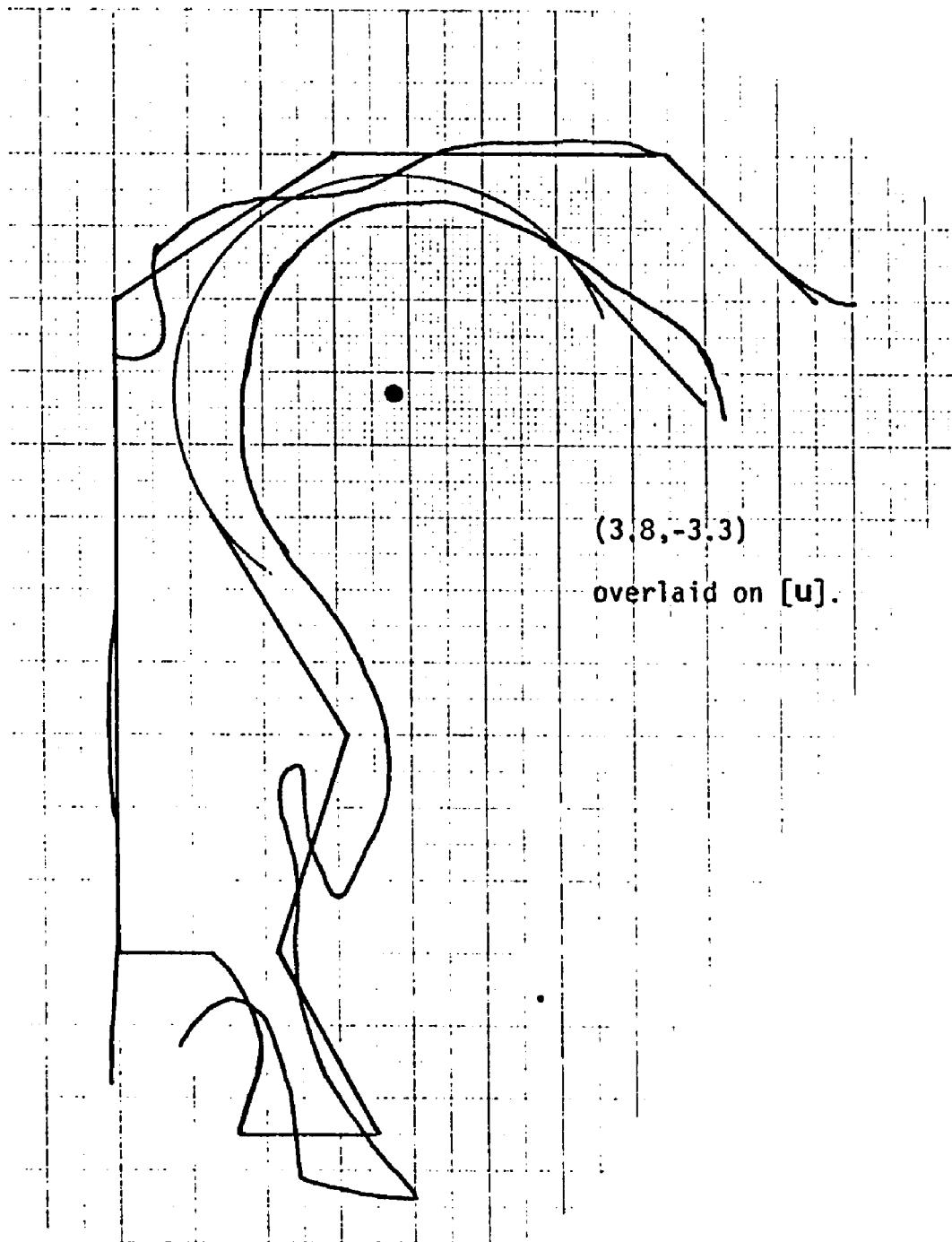
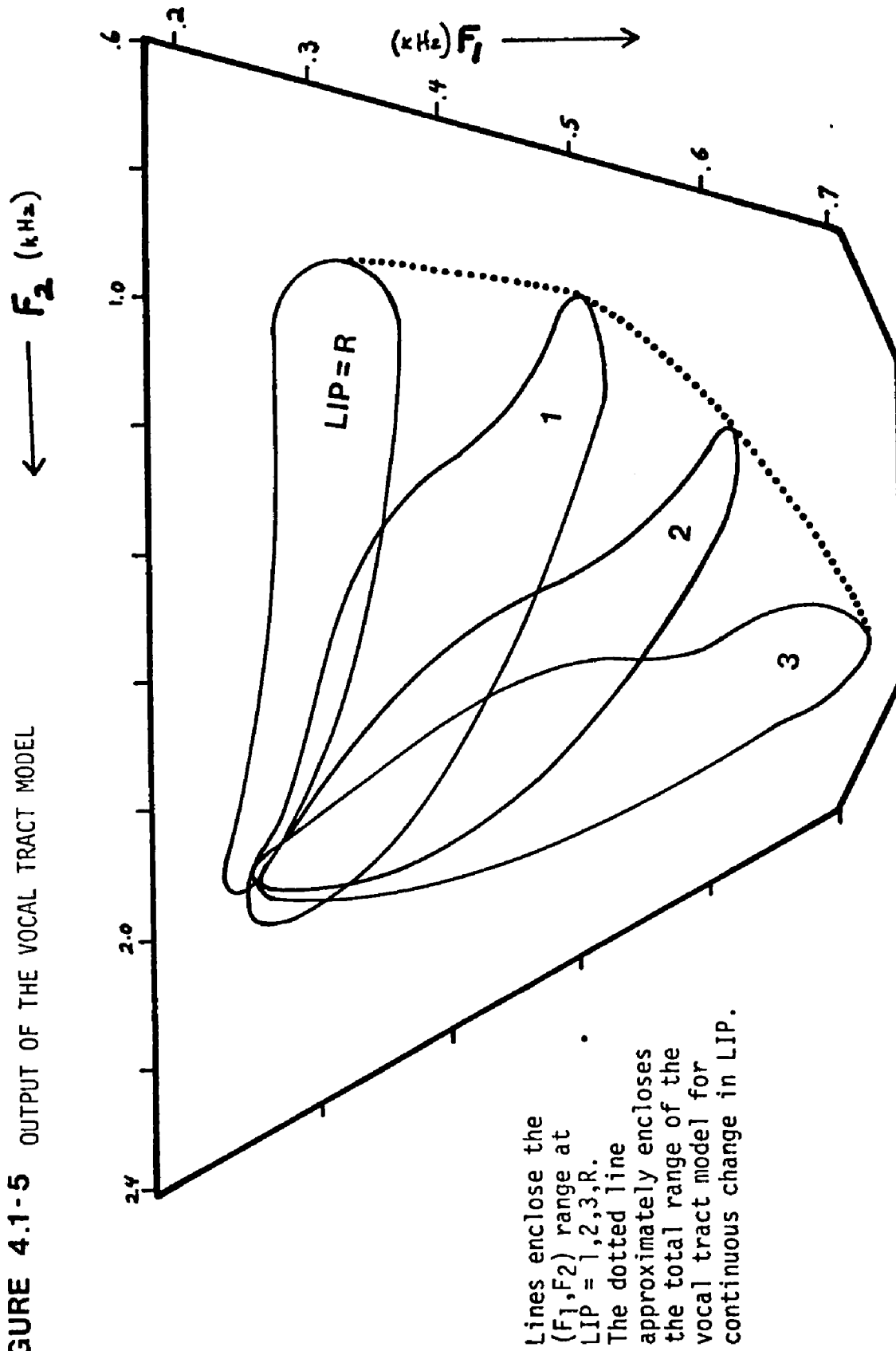
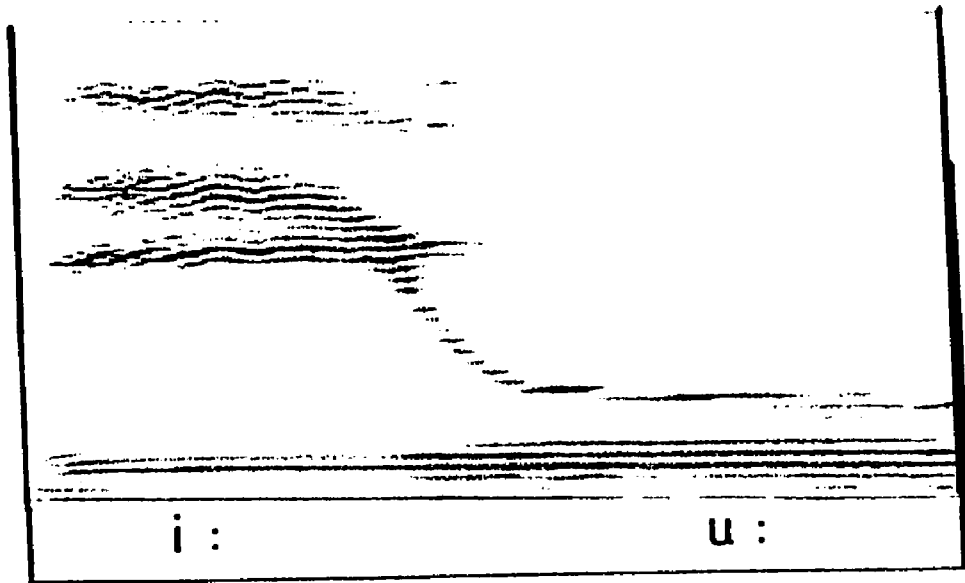


FIGURE 4.1-4 (cont.)

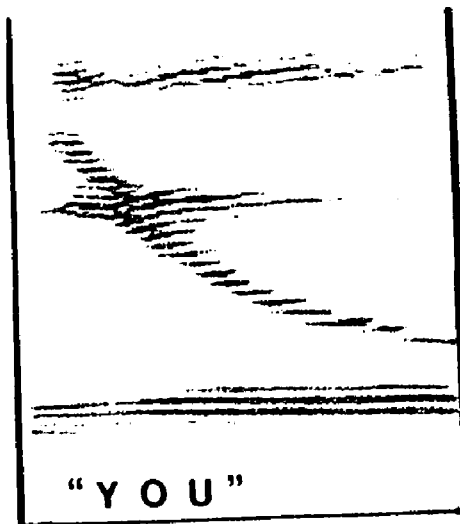
FIGURE 4.1-5 OUTPUT OF THE VOCAL TRACT MODEL



(a) Spectrogram of
[i:u:] pronounced
by JH



(b) Spectrogram of "you"
pronounced by JH



(c) Formant movements
of JH's [i:u:] seen
in F_2 by F_3 plane

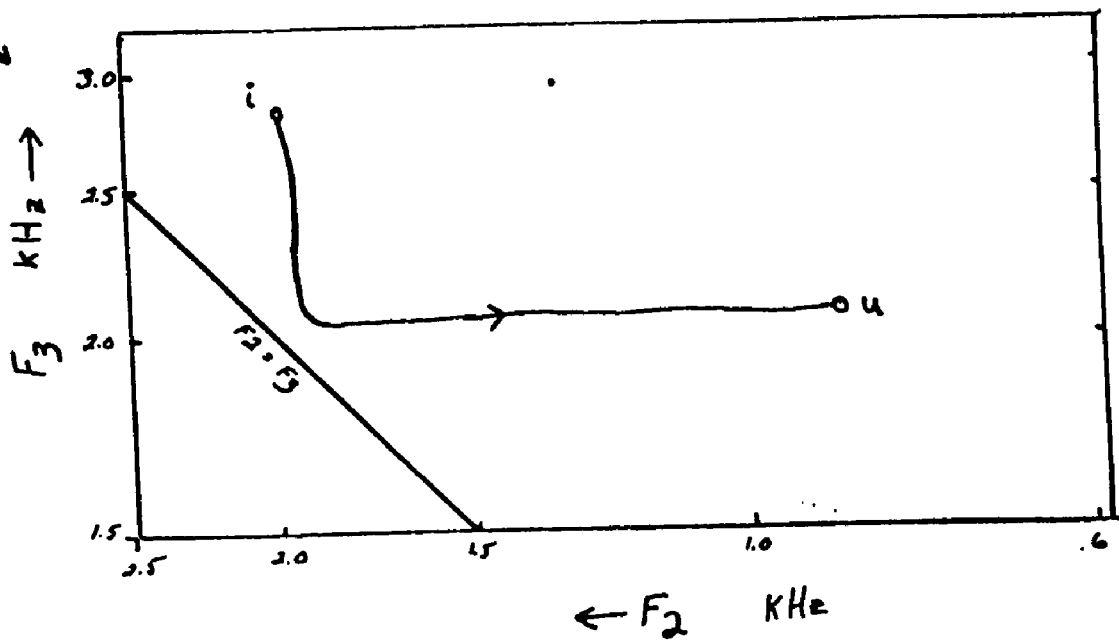


FIGURE 4.1-6

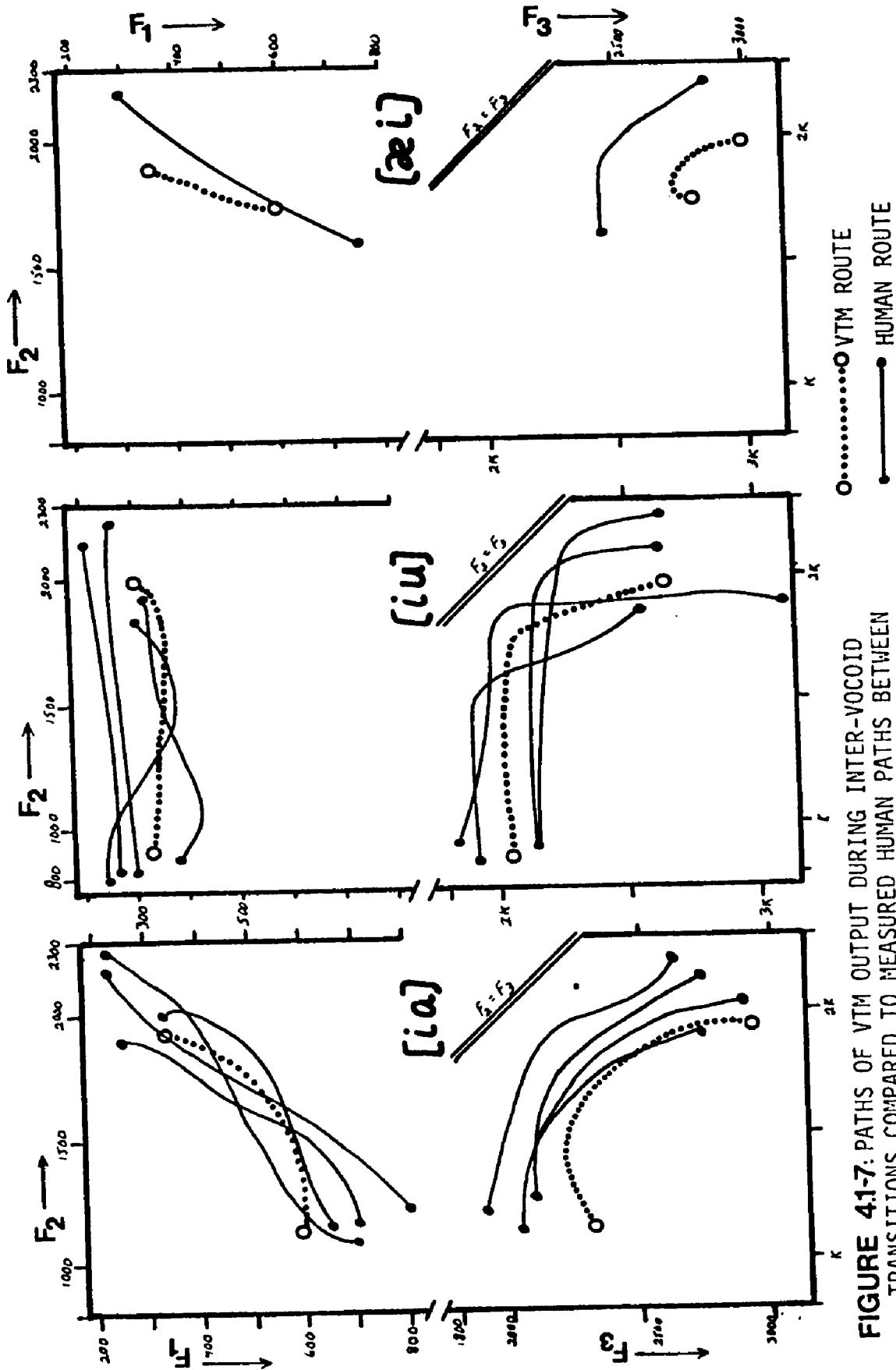


FIGURE 4.1-7: PATHS OF VTM OUTPUT DURING INTER-VOCOID TRANSITIONS COMPARED TO MEASURED HUMAN PATHS BETWEEN SIMILAR TARGETS.

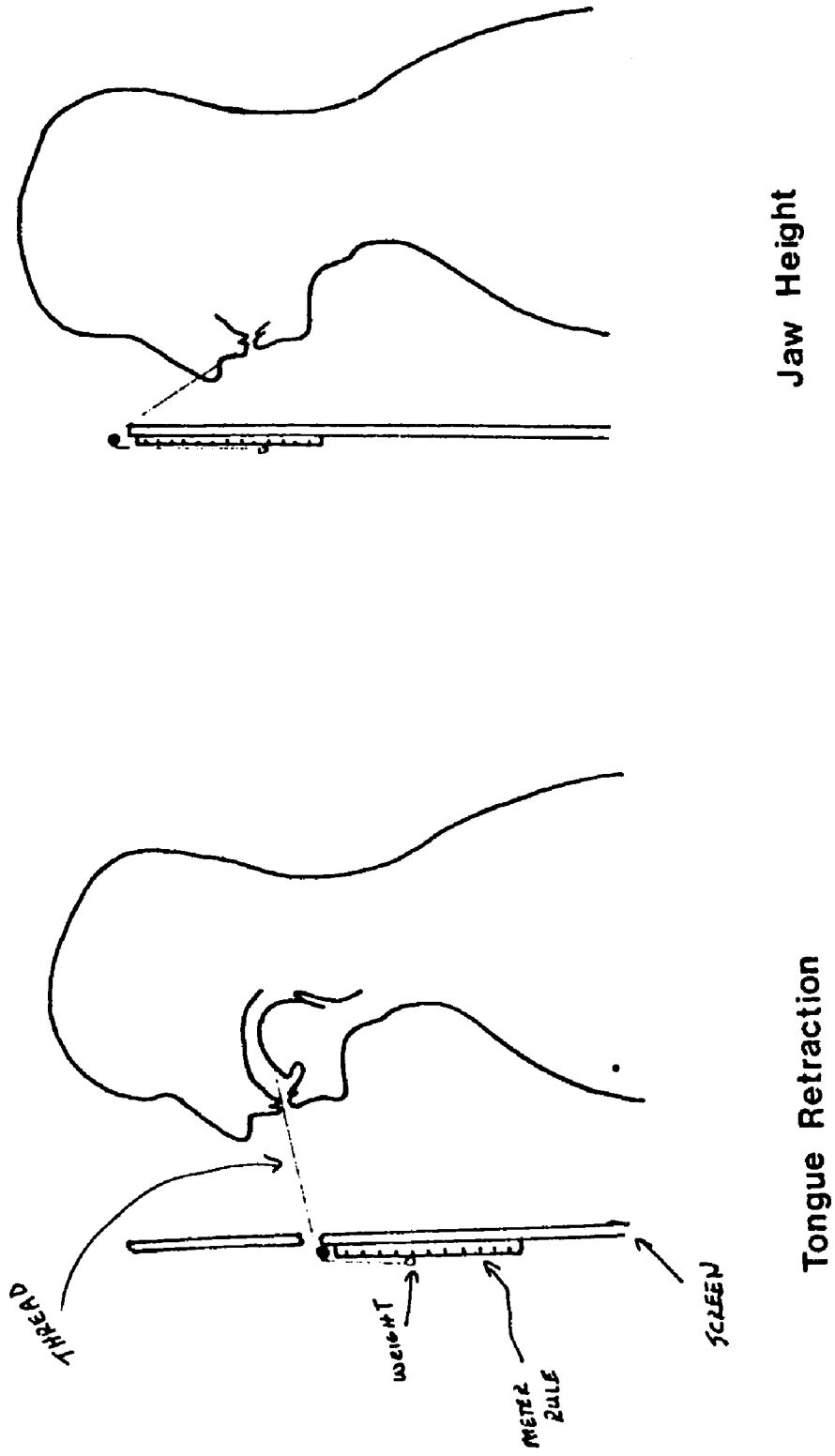


FIGURE 4.2-1 : Apparatus for Articulator Proprioception Experiment

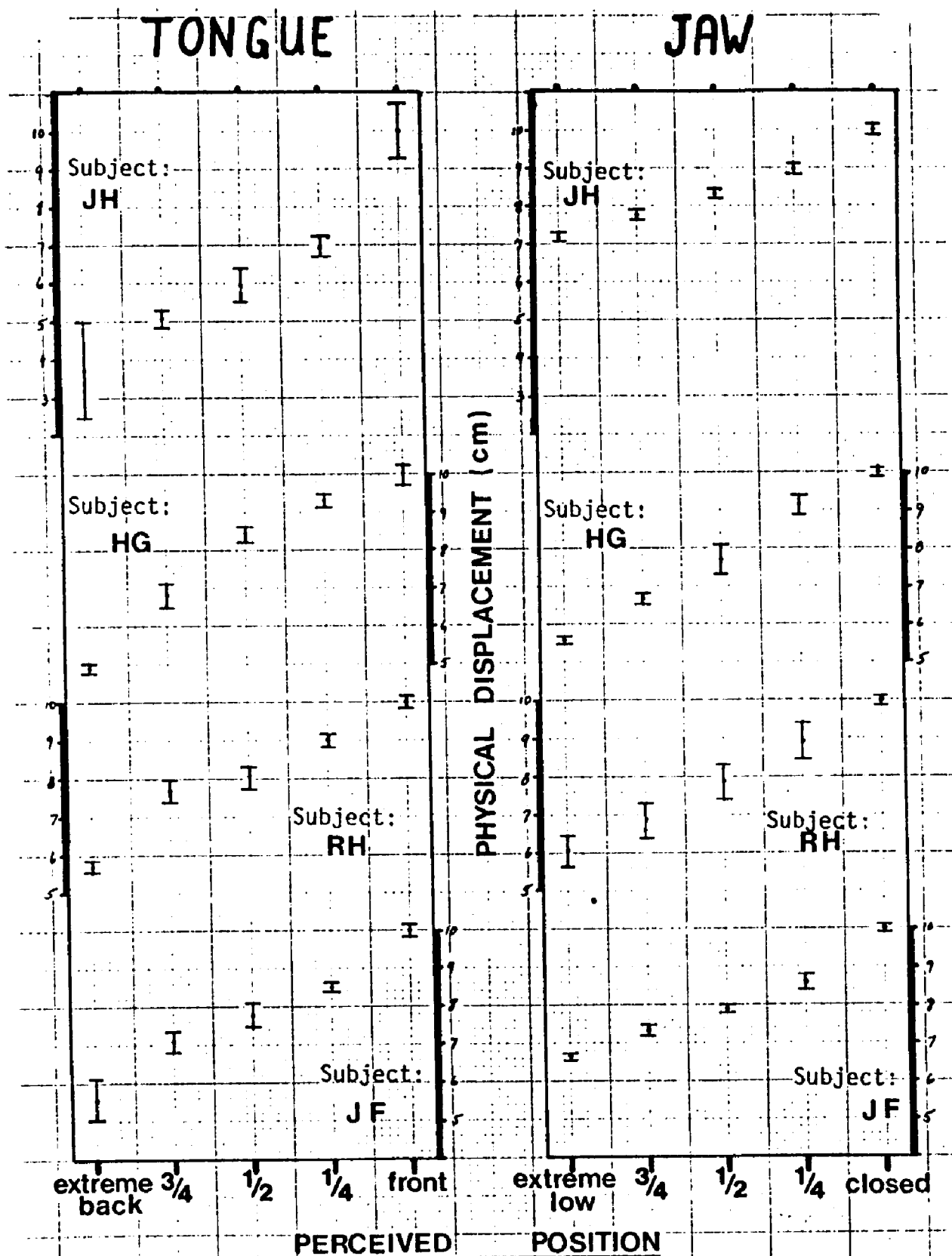


FIGURE 4.2-2 PROPRIOCEPTION OF ARTICULATOR POSITION

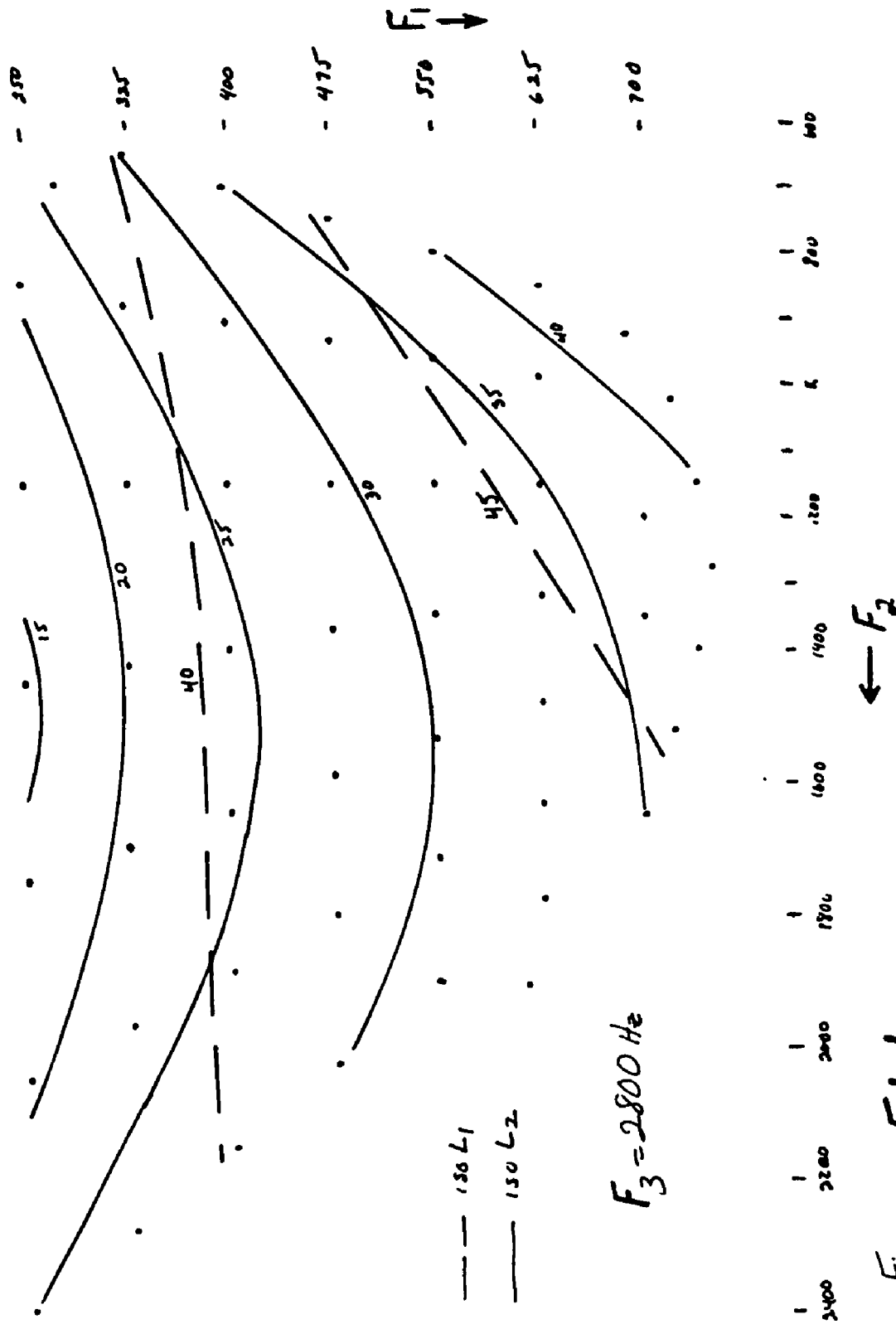


Figure 5.1-1

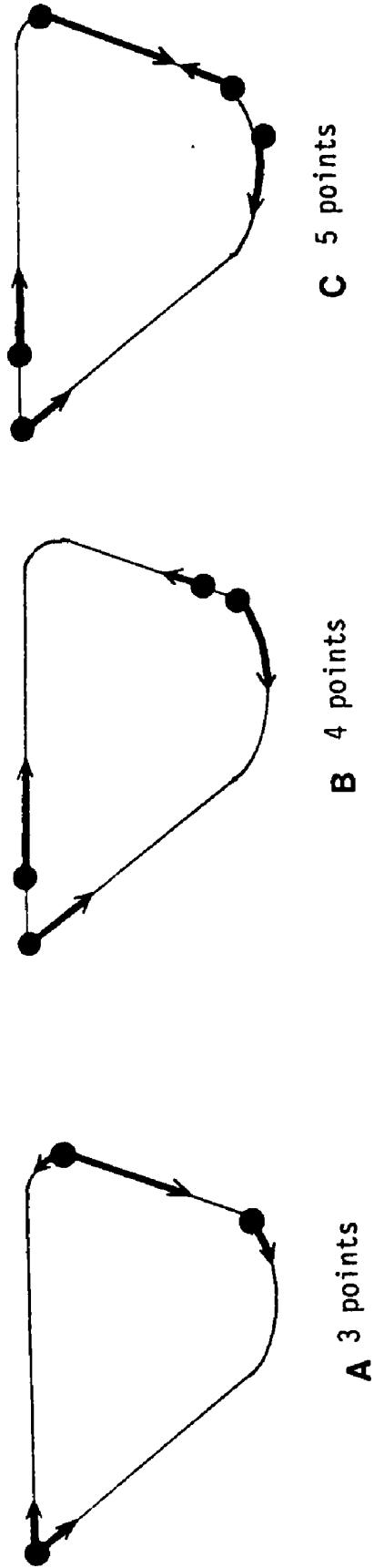


FIGURE 5.1-2: MAXIMAL 3, 4, 5 POINT CONFIGURATIONS.

Arrows show range of movement for individual points (others fixed at maximal) producing less than 1% reduction in aggregate distance.

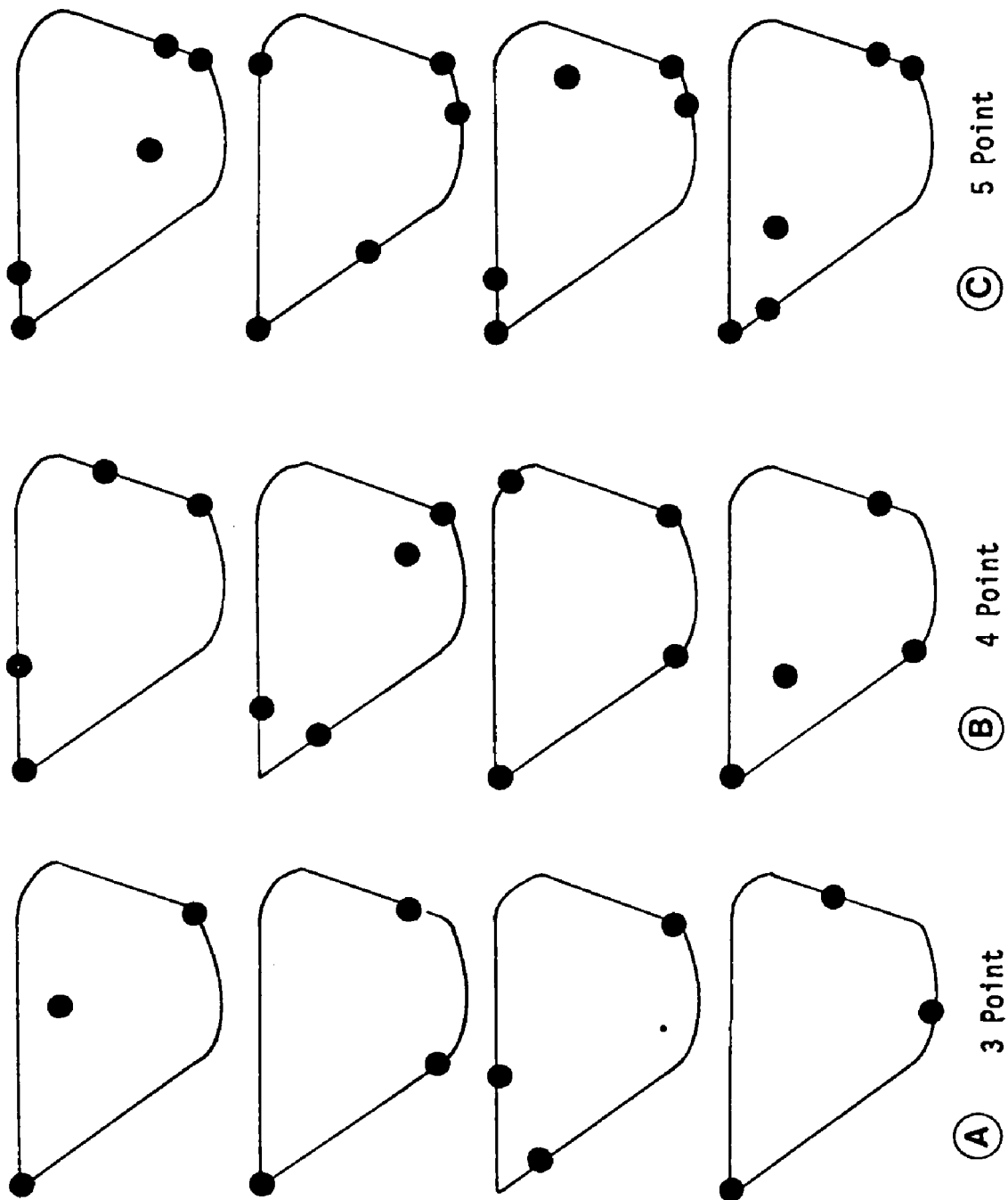
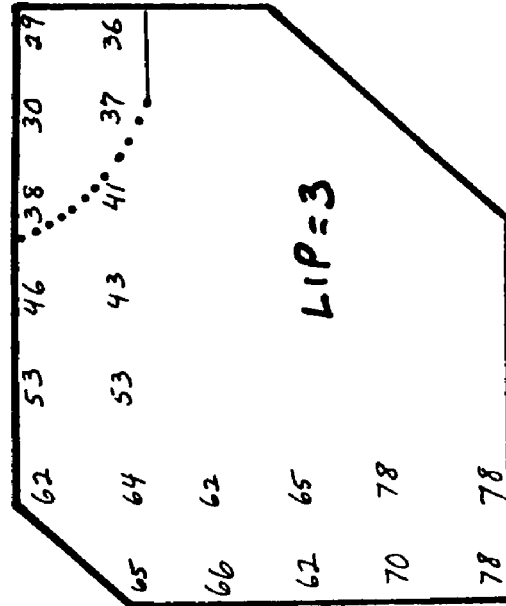
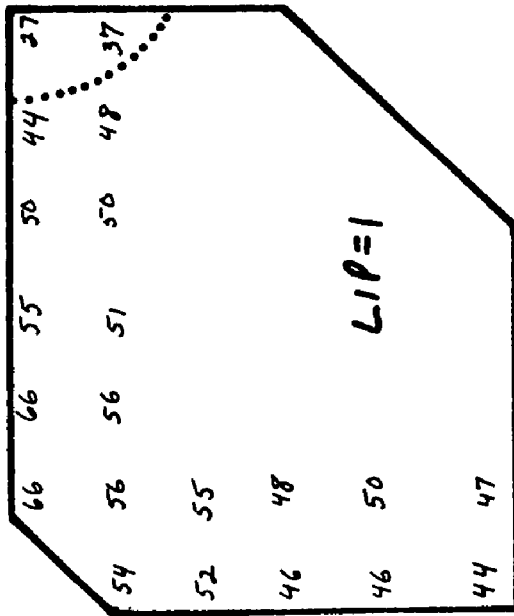
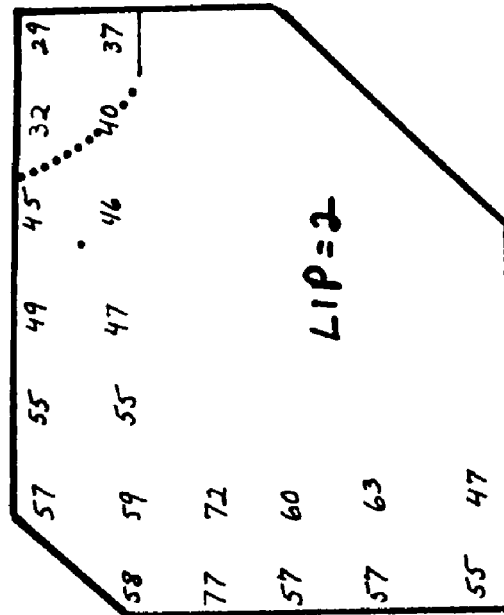
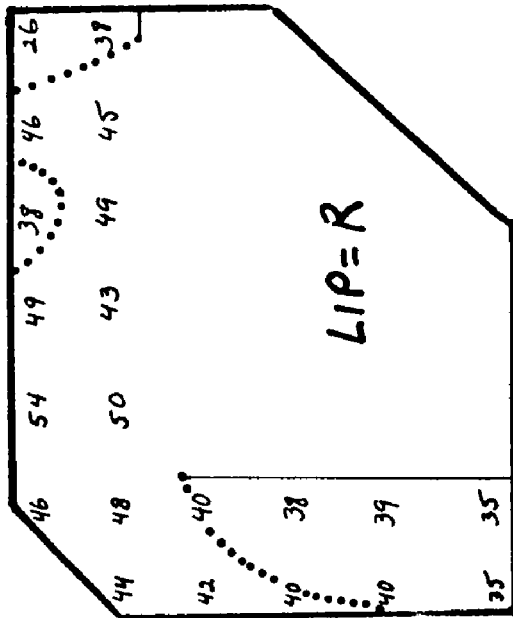


FIGURE 5.1-3 SELECTED 3, 4, AND 5 POINT CONFIGURATIONS
WITHIN 5% OF MAXIMAL



LESS THAN
40

FIGURE 5.2-1: HAND CALCULATED
SCALAR GRADIENT FOR BACK AND HIGH
REGIONS OF TX, TY.
See Figure 4.1-3 for coordinates.



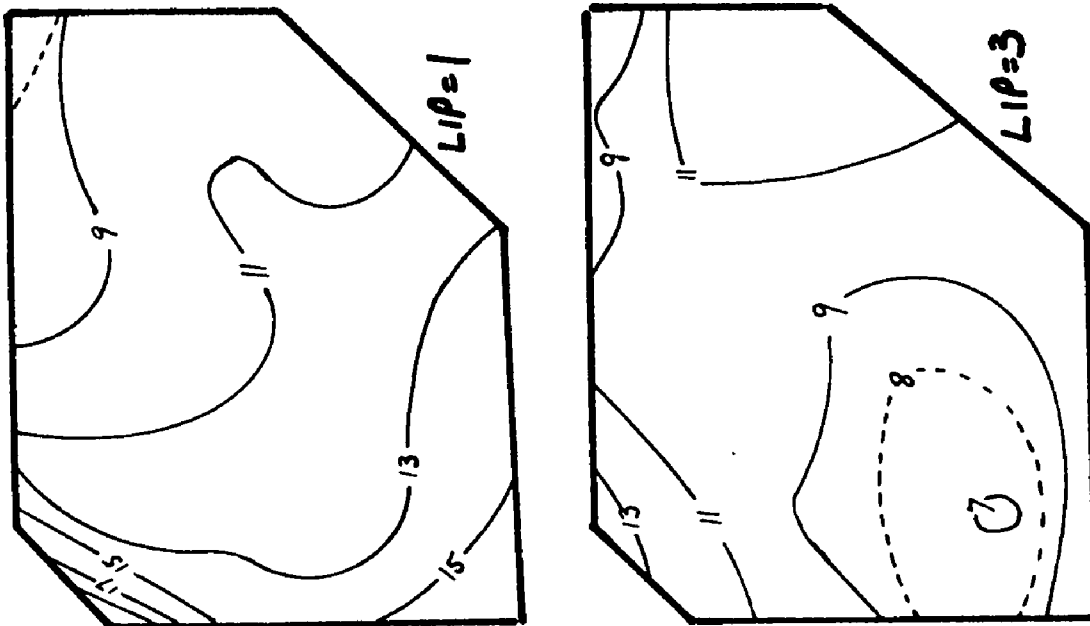
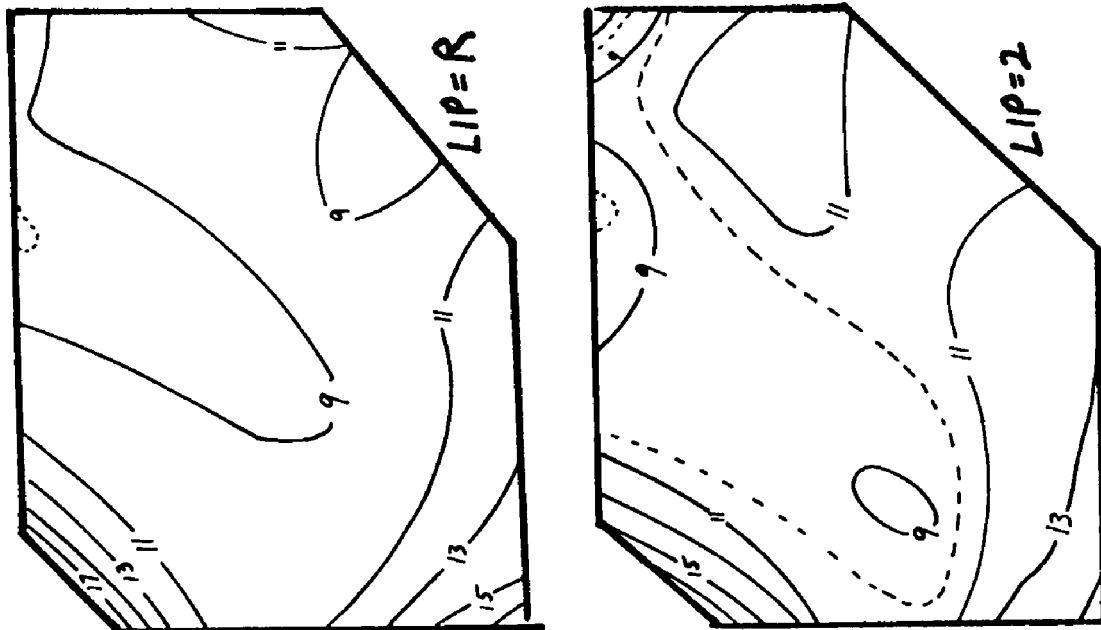


FIGURE 5.2-2: CONTOURS OF EQUAL COMPUTED GRADIENCE.

LOWER NUMBERS INDICATE THE MORE STABLE
OR "QUANTAL" REGIONS. See Figure 4.1-3
for coordinates.



REFERENCES

- Abboud, P., Bezirgan, N., Erwin, W., Khouri, M., McCarus, E., and Rammuny, R. (1968). Elementary Modern Standard Arabic. Inter-University Committee for Near Eastern Languages, Ann Arbor.
- Barclay, J. (1972). "Noncategorical perception of a voiced stop." Perception and Psychophysics, Vol. 11, p. 269-272.
- Beals, R., Krantz, D., and Tversky, A. (1968). "Foundations of multidimensional scaling." Psychological Review, Vol. 75, p. 127-142.
- Bishop, M., Ringel, R., and House, A. (1973). "Orosensory perception, speech production, and deafness." Journal of Speech and Hearing Research, Vol. 16, p. 257-266.
- Carlson, R., Granström, B., and Fant, G. (1970). "Some studies concerning the perception of isolated vowels." KTH STL-QPSR 2-3/1970, p. 19-35.
- Catford, J. (1964). "Phonation types." In D. Abercrombie et al. (Eds.), In Honour of Daniel Jones. Longmans, London.
- Catford, J. (1968). "The articulatory possibilities of man." In B. Malmberg (Ed.), Manual of Phonetics. North-Holland Pub. Co., Amsterdam.
- Catford, J. (1974). "'Natural' sound changes." In A. Bruck (Ed.), Papers from the Parasession on Natural Phonology. Chicago Ling. Soc., Chicago, p. 21-29.
- Chiba, T., and Kajiyama, M. (1941). The Vowel. Tokyo-Kaiseikan Pub. Co., Tokyo.
- Chomsky, N., and Halle, M. (1968). The Sound Pattern of English. Harper Row, New York.
- Coker, C., Umeda, N., and Browman, C. (1973). "Automatic synthesis from ordinary English Text." In J. Flanagan and L. Rabiner (Eds.), Speech Synthesis. Dowden, Hutchinson, & Ross, Stroudsburg, Pa.
- Delattre, P., Liberman, A., Cooper, F., and Gerstman, L. (1952). "An experimental study of the acoustic determinants of vowel color." Word, Vol. 8, p. 195-210.
- Dunn, H. (1961). "Methods of measuring vowel formant bandwidths." Journal of the Acoustical Society of America, Vol. 33, p. 1737-1746.
- Eriksen, C., and Hake, H. (1955). "Multidimensional stimulus differences and accuracy of discrimination." Journal of Experimental Psychology, Vol. 50, p. 153-160.

- Erwin, W. (1963). A Short Reference Grammar of Iraqi Arabic. Georgetown U. Press, Washington, D.C.
- Fairbanks, G., and Grubb, P. (1961). "A psychophysical investigation of vowel formants." Journal of Speech and Hearing Research, Vol. 4., p. 203-219.
- Fant, G. (1966). "A note on vocal tract size factors and non-uniform F-pattern scalings." KTH STL-QPSR 4/1966, Stockholm. (reprinted in Fant, 1973).
- Fant, G. (1970). Acoustic Theory of Speech Production. Mouton, The Hague.
- Fant, G. (1972). "Vocal tract wall effects, losses, and resonance bandwidths." KTH STL-QPSR 2-3/1972, p. 28-52.
- Fant, G. (1973). Speech Sounds and Features. MIT Press, Cambridge.
- Fant, G. (1976). "Vocal tract area and length perturbations." KTH STL-QPSR 4/1975, p. 1-14.
- Fischer-Jørgensen, E. (1968). "Perceptual dimensions of vowels." Z. Phonet. Spr. Komm, Vol. 21, p. 94-98.
- Flanagan, J. (1955a). "A difference limen for vowel formant frequency." Journal of the Acoustical Society of America, Vol. 27, p. 613-617. (reprinted in Lehiste, 1967).
- Flanagan, J. (1955b). "Difference limen for the intensity of a vowel sound." Journal of the Acoustical Society of America, Vol. 27, p. 1223-1225.
- Flanagan, J. (1957a). "Estimates of the maximum precision necessary in quantizing certain 'dimensions' of vowel sounds." Journal of the Acoustical Society of America, Vol. 29, p. 553-554. (reprinted in Lehiste, 1967).
- Flanagan, J. (1957b). "Difference limen for formant amplitude." Journal of Speech and Hearing Disorders, Vol. 22, p. 205-212. (reprinted in Lehiste, 1967).
- Flanagan, J. (1965). "Recent studies in speech research at Bell Telephone Laboratories (II)." Proceedings of the Fifth International Congress on Acoustics, Liege, Belgium.
- Flanagan, J. (1972). Speech Analysis, Synthesis, and Perception (2nd ed.). Springer, Berlin.
- Fromkin, V. (1968). "The what and why of intrinsicness." Working Papers in Phonetics, #10. UCLA, p. 156-168.
- Fudge, E. (1973). "On the notion of 'Universal Phonetic Framework.'" In E. Fudge (Ed.), Phonology. Penguin, Baltimore.
- Fujimura, O. (1967). "On the second spectral peak of front vowels." Language and Speech, Vol. 10, p. 181-193.

- Fujimura, O., and Lindquist, J. (1971). "Sweep-tone measurements of the vocal tract characteristics." Journal of the Acoustical Society of America, Vol. 49, p. 541-558.
- Halle, M., and Stevens, K. (1971). "A note on laryngeal features." Quarterly Progress Report No. 101, p. 198-213, Res. Lab. Elec., MIT.
- Hanson, G. (1967). "Dimensions in speech sound perception." Ericsson Technics, No. 1.
- Heinz, J., and Stevens, K. (1965). "On the relation between lateral cineradiographs, area functions, and acoustic spectra of speech." Paper A44 in Proceedings of the Fifth International Congress on Acoustics. American Institute of Physics, New York.
- Hjelmslev, L. (1936). "On the principles of phonematics." Proceedings of the Second International Congress of Phonetic Sciences, p. 49-54.
- Hockett, C. (1955). A Manual of Phonology. Memoir 11 of IJAL. Waverly Press, Baltimore.
- Holden, K. (1972). "Loan words and phonological systems." Unpublished dissertation, University of Texas.
- Jakobson, R. (1968). Child Language Aphasia and Phonological Universals. Mouton, The Hague.
- Jakobson, R., Fant, G., and Halle, M. (1963). Preliminaries to Speech Analysis. MIT Press, Cambridge.
- Jakobson, R., and Halle, M. (1968). "Phonetics in relation to phonology." In B. Malmberg (Ed.), The Manual of Phonetics. North-Holland Pub. Co., Amsterdam.
- Jonasson, J. (1971). "Perceptual similarity and articulatory reinterpretation as a source of phonological innovation." KTH STL-QPSR 1/1971, p. 30-42.
- Jones, D. (1932). An Outline of English Phonetics. Heffer, Cambridge.
- Jones, D. (1967). The Phoneme. Heffer, Cambridge.
- Kahn, M. (1975). "Arabic emphatics: The evidence for cultural determinants of phonetic sex-typing." Phonetica, Vol. 31, p. 38-50.
- Kent, R. (1972). "Some considerations in the cinefluorographic analysis of tongue movements during speech." Phonetica, Vol. 26, p. 16-32.
- Krantz, D., Luce, R., Suppes, P., and Tversky, A. (1971). Foundations of Measurement. Academic Press, New York.

- Krantz, D., and Tversky, A. (1971). "Conjoint measurement analysis of composition rules in psychology." Psychological Review, Vol. 78, p. 157-169.
- Krantz, D., and Tversky, A. (1975). "Similarity of rectangles." Journal of Mathematical Psychology, Vol. 12, p. 4-34.
- Labov, W., Yaeger, M., and Steiner, R. (1972). A Quantitative Study of Sound Change in Progress (2 volumes). U.S. Regional Survey, Philadelphia.
- Ladefoged, P. (1964). A Phonetic Study of West African Languages. Cambridge, U. Press, Cambridge.
- Ladefoged, P. (1967). Three Areas of Experimental Phonetics. Oxford U. Press, London.
- Ladefoged, P. (1971). Preliminaries to Linguistic Phonetics. U. Chicago Press, Chicago.
- Ladefoged, P. (1975). A Course in Phonetics. Harcourt Brace, New York.
- Ladefoged, P., Anthony, J., and Riley, C. (1971). "Direct measurement of the vocal tract." Working Papers in Phonetics #19. UCLA, p. 4-13.
- Lambton, A. (1971). Persian Grammar. Cambridge U. Press, Cambridge.
- LaPointe, L., and Wertz, R. (1974). "Oral-movement abilities and articulatory characteristics of brain-injured adults." Perceptual and Motor Skills, Vol. 39, p. 39-46.
- Lehiste, I. (Ed.). (1967). Readings in Acoustic Phonetics. MIT Press, Cambridge.
- Lehiste, I. (1970). Suprasegmentals. MIT Press, Cambridge.
- Liberman, A. (1970). "The grammars of speech and language." Cognitive Psychology, Vol. 1, p. 301-323.
- Liljenkrants, J. (1971). "A Fourier series description of the tongue profile." KTH STL-QPSR 4/1971, p. 9-18.
- Liljenkrants, J., and Lindblom, B. (1972). "Numerical simulation of vowel quality systems." Language, Vol. 48, p. 839-862.
- Lindblom, B. (1963). "On vowel reduction." Report #29, Speech Trans. Lab., Royal Inst. of Tech., Stockholm.
- Lindblom, B. (1972). "Phonetics and the description of language." In A. Rigault (Ed.), Proceedings of the Seventh International Congress of Phonetic Sciences. Mouton, The Hague.
- Lindblom, B., and Sundberg, J. (1969). "A quantitative model of vowel production and the distinctive features of Swedish vowels." KTH STL-QPSR 1/1969, p. 14-32.

- Lindblom, B., and Sundberg, J. (1971). "Acoustical consequences of lip, tongue, jaw, and larynx movement." Journal of the Acoustical Society of America, Vol. 50, p. 1166-1179.
- Maeda, S. (1972). "On the conversion of vocal tract x-ray data into formant frequencies." Bell Laboratories, Murray Hill, N.J. Abstracted in the Journal of the Acoustical Society of America, Vol. 51, p. 88.
- Mermelstein, P. (1967). "Determination of the vocal-tract shape from measured formant frequencies." Journal of the Acoustical Society of America, Vol. 41, p. 1283-1294.
- Mermelstein, P. (1973). "Articulatory model for the study of speech production." Journal of the Acoustical Society of America, Vol. 53, p. 1070-1082.
- Miller, G. (1956). "The magical number seven, plus or minus two." Psychological Review, Vol. 63, p. 81-96.
- Miller, R. (1953). "Auditory tests with synthetic vowels." Journal of the Acoustical Society of America, Vol. 25, p. 114-121.
- Öhman, S. (1967). "Numerical model of coarticulation." Journal of the Acoustical Society of America, Vol. 41, p. 310-320.
- Paddock, H. (1970). "The major pitch features of vocalic quality." Lingua, Vol. 25, p. 142-151.
- Paget, R. (1930). Human Speech. Kegan Paul, London.
- Paris, C. (1974). Systeme Phonologique du Besney. Publication LXIX; Societe de Linguistique de Paris, Paris.
- Perkell, J. (1969). Physiology of Speech Production. MIT Press, Cambridge.
- Perkell, J. (1974). "A physiologically-oriented model of tongue activity in speech production." Ph.D. Thesis, MIT.
- Peterson, G. (1961). "Parameters of vowel quality." Journal of Speech and Hearing Research, Vol. 4, p. 10-29.
- Peterson, G., and Barney, H. (1952). "Control methods used in a study of vowels." Journal of the Acoustical Society of America, Vol. 24, p. 175-184.
- Pike, K. (1943). Phonetics. U. of Michigan Press, Ann Arbor.
- Pike, K. (1947). Phonemics. U. of Michigan Press, Ann Arbor.
- Pike, K. (1967). Language in Relation to a Unified Theory of the Structure of Human Behavior. Mouton, The Hague.
- Pisoni, D. (1973a). "The role of auditory short-term memory in vowel perception." Haskins Laboratories: SR-34, p. 89-117.

- Pisoni, D. (1973b). "Auditory and phonetic memory codes in the discrimination of consonants and vowels." Perception and Psychophysics, Vol. 13, p. 253-260.
- Pollack, I. (1953). "The information of elementary auditory displays. II." Journal of the Acoustical Society of America, Vol. 25, p. 765-769.
- Pollack, I., and Ficks, L. (1954). "Information of elementary multidimensional auditory displays." Journal of the Acoustical Society of America, Vol. 26, p. 155-158.
- Pols, L., Kamp, L., and Plomp, R. (1969). "Perceptual and physical space for vowel sounds." Journal of the Acoustical Society of America, Vol. 46, p. 458-467.
- Pols, L., Tromp, H., and Plomp, R. (1973). "Frequency analysis of Dutch vowels from 50 male speakers." Journal of the Acoustical Society of America, Vol. 53, p. 1093-1101.
- Russel, G. (1928). The Vowel. Ohio State U. Press, Columbus.
- de Saussure, F. (1959). Course in General Linguistics. McGraw-Hill, New York.
- Schroeder, M. (1967). "Determination of the geometry of the human vocal tract by acoustic measurements." Journal of the Acoustical Society of America, Vol. 41, p. 1002-1010.
- Sedlack, P. (1969). "Typological considerations of vowel quality systems." Working Papers on Language Universals, no. 1. Stanford U.
- Staalsen, P. (1966). "The phonemes of Iatmul." Pacific Linguistics, Series A, occasional paper no. 7, p. 69-76.
- Stampe, D. (1972). "How I spent my summer vacation." Unpublished Ph.D. Dissertation, U. Chicago.
- Stetson, R. (1951). Motor Phonetics. North-Holland Pub. Co., Amsterdam.
- Stevens, K. (1972). "The quantal nature of speech." In P. Denes and E. David (Eds.), Human Communication, A Unified View. McGraw-Hill, New York.
- Stevens, K., and House, A. (1955). "Development of a quantitative description of vowel articulation." Journal of the Acoustical Society of America, Vol. 27, p. 484-493. (reprinted in Lehiste, 1967).
- Stevens, K., and House, A. (1961). "An acoustical theory of vowel production and some of its implications." Journal of Speech and Hearing Research, Vol. 4, p. 303-320. (reprinted in Lehiste, 1967).

- Stevens, K., and House, A. (1972). "Speech perception." In J. Tobias (Ed.), Foundations of Modern Auditory Theory. Academic Press, New York.
- Sundberg, J. (1969). "On the problem of obtaining area functions from lateral x-ray pictures of the vocal tract." KTH STL-QPSR 1/1969, p. 43-46.
- Terbeek, D., and Harshman, R. (1971). "Cross-language differences in the perception of natural vowel sounds." Working Papers in Phonetics #19. UCLA, p. 26-38.
- Terbeek, D., and Harshman, R. (1972). "Is vowel perception non-Euclidean." Working Papers in Phonetics #22. UCLA, p. 13-29.
- Trubetzkoy, N. (1969). Principles of Phonology. U. of California Press, Berkeley.
- Tversky, A., and Krantz, D. (1970). "The dimensional representation and the metric structure of similarity data." Journal of Mathematical Psychology, Vol. 7, p. 572-596.
- Wickelgren, W. (1965). "Distinctive features and errors in short-term memory for English vowels." Journal of the Acoustical Society of America, Vol. 38, p. 583-588.