iu linguistics club

phonetic feature systems for vowels

Terrance Michael Nearey

T. M. Neary

PHONETIC FEATURE SYSTEMS

FOR VOWELS

by

Terrance Michael Nearey

UNIVERSITY OF ALBERTA

Reproduced by the
Indiana University Linguistics Club
310 Lindley Hall
Bloomington, Indiana 47401
December, 1978

notes on published paper
$146

p. 2.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS

This work would not have been possible without the patient support of my wife, Beatrice, and my daughter, Laura. It is to them this work is warmly dedicated.

I wish also to thank the members of my committee for general guidance of this work as a whole. In addition, each has made special contributions that deserve mention. Dr. Philip Lieberman provided the original suggestion that led to this work: the inadequacy of the articulatory features for vowels. Dr. Thomas Gay aided greatly in the planning of the cineradiographic experiments and guided its analysis. Suggestions and comments made by Dr. Arthur S. Abramson have greatly influenced the theoretical discussion of Chapter I.

Apart from my committee, there are several other persons who have been instrumental in the conception and execution of this project. Dr. I. G. Mattingly has provided helpful suggestions and important critical comments, particularly with regard to the analysis and synthesis of speaker differences. He also kindly supplied me with the individual subjects' data from Peterson and Barney (1952). I would also like to thank my fellow student Robert F. Port. The analysis of speaker differences presented in Chapter IV owes its ultimate inspiration to discussions with him.

I wish to thank Dr. Samuel Zahl of the Department of Statistics, the University of Connecticut, for his suggestions and guidance in the statistical analyses of Chapter IV. I also thank Dr. Norman Draper of the Department of Statistics, the University of Wisconsin, for his suggestion of the application of the method of Box and Cox.

Members of the staff of Haskins Laboratories have aided various aspects of this enterprise. I thank Dr. A. M. Liberman, President of Haskins Laboratories, for the use of facilities in the analysis of the acoustic and cineradiographic data. Discussions with him have also been helpful in the analysis of the issues in the perceptual experiment of Chapter III. Dr. T. Ushijima prepared the subjects for the radiographic experiments. I thank Dr. F. S. Cooper and Dr. George N. Sholes for serving as subjects in the radiographic experiment. I am also indebted to Drs. Katherine Harris, Donald Shankweiler, Paul Mermelstein and Michael Studdert-Kennedy for comments and suggestions at various stages of this research.

Thanks are also due to Katherine Watson who prepared many of the drawings for figures, and to Mrs. Reta Koslo for the typing of the tables and captions to the figures. Computer facilities of the

Department of Linguistics and the Computer Center at the University of Connecticut were used extensively for stimulus generation and data analysis. The text was produced by the FMT system of the Computer Center of the University of Alberta.

Parts of this research were supported by grants from the National Institute of Dental Research (through Haskins Laboratories) and the Research Foundation of the University of Connecticut.

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

The principal claim made in this work is that there is strong evidence that phonetic features specifying vowels are more directly related to acoustic rather than articulatory parameters.

In Chapter I, it is argued that certain traditional assumptions about the nature of phonetic representations imply that phonetic features are derivable as language-independent functions of physical parameters. It is further argued that the acoustic signal contains all the information necessary for the recovery of phonetic representations. While it is possible that this recovery takes place only through mediation of an articulatory representation, the fact that different articulatory configurations may produce the same output raises serious theoretical problems for articulatorily-oriented feature systems. The structure and history of traditional accounts of vowel quality is discussed with emphasis on the question of auditory versus articulatory description.

In Chapter II, radiographic evidence from the literature bearing on the validity of traditional articulatory descriptions is discussed and new X-ray evidence is presented. Here it is argued that virtually every radiographic study involving more than one subject has indicated 1) a general relationship structure substantially different from traditional descriptions 2) intersubject variability that is evidently unsystematic in that there appears to be no general way of separating speaker-dependent from phone-dependent variation.

In the last two chapters, the problem of cross-speaker variation in acoustic parameters of vowels is examined. While acoustic analyses have indicated relatively mild variation in the acoustic parameters of subjects of the same sex-age class, the variation can be extreme when the formant values of adult males are compared with those of females or children.

Chapter III discusses several hypotheses that imply that vowel quality features are derivable as functions of the RELATIONSHIPS in the formant frequencies of a single speaker's vowel system. The question of the relevance of formant relationships to vowel perception is also examined. One important set of relative-formant hypotheses are what may be termed constant ratio hypotheses. Such models claim that formant frequencies from different subjects are relatable by means of one or more multiplicative scale factors. These models imply that the information contained in a single vowel-point of known phonetic quality should be sufficient to determine a speaker's entire vowel system. The results of a

perceptual experiment designed to test this prediction are presented in Chapter III. They indicate that a change in formant frequencies of a single context (or "carrier") vowel is sufficient to change the categorization of a two-dimensional (F1-F2) continuum of stimuli in a manner qualitatively compatible with a point-normalization model.

In Chapter IV, several relative-formant normalization procedures are tested on empirical data in the form of formant measurements from real speech provided in the literature. All the methods investigated result in generally high rates of correct identification. The point-normalization hypotheses of the type generally compatible with the results of the perceptual experiment presented in Chapter III rival (though do not surpass) more complex models in their phonetic resolving power (their ability to separate vowels of differing phonetic qualities). Some arguments against the notion of constant ratios in the formant frequencies of different speakers are discussed. While it appears that the extent of any deviations from constant ratios is limited, an attempt is made to arrive at an improved point normalization model.

CHAPTER I

PHONETIC FEATURES AND TRADITIONAL FEATURES FOR VOWEL QUALITY

1.1 The Nature of Phonetic Features

Introduction

It is argued below that traditional assumptions about the nature of phonetic representation and its relationship to physical events leads naturally to the conclusion that phonetic features are specified as functions of physical parameters. It is further argued that features are recoverable from the acoustic signal, whether or not some level of articulatory representation mediates this recovery. It is important to note that phonetic features are not here assumed a priori to be more closely associated with either acoustic or articulatory parameters. Henceforth, the term "feature" will be used to signify the basic unit of the structure of phonetic representation. The term "parameter" will be used to refer to aspects of physical events. Perhaps the most fundamental task facing experimental phonetics is that of specifying the mapping that exists between physical parameters and phonetic representations.

At least three logically distinct objects may be referred to by the term "phonetic representation": 1) the transcription and allied feature analyses of trained phoneticians; 2) the output of a phonology of a grammar; 3) a psychological object that is at the level of linguistic representation nearest to the articulatory-acoustic events that constitute a speech act. The term "transcription" will be reserved specifically for transcription by trained listeners. It will generally be assumed that the three logically distinct objects are closely related to one another. Unless otherwise noted, the discussion below will be concerned with the most readily accessible form of phonetic representation, transcription. However, on occasion specific reference will be made to the psychological object of item three above and its relation to transcription.

The internal structure of phonetic representation

Universal segments. -- The term "universal character" used by many 17th century phoneticians implies two principles that underlie virtually all phonetic work to the present day. The first of these is that phonetic representations are universal; that is,

language-independent. It will be argued below that this implies the existence of a regular, language-independent relationship between sets of physical parameters and phonetic representations.

The second principle implicit in the notion "universal character" is that phonetic events are representable by (alphabetic) characters, or segments. Segmentation implies that all utterances with different phonetic representations are not arbitrarily and monolithically different from each other, but rather that they can be analysed into partly identical strings. For a summary of empirical evidence supporting segmental structure the reader is referred to the work of Studdert-Kennedy (1974). The assumption of universality will be discussed below.

Features.-- Most phonetic analyses since the 17th century have also assumed that segments are not arbitrarily and monolithically distinct. Segments have been further analysed into features by which each phonetic segment 1) enters into relationships of partial identity with some other segments; and 2) enters into relationships of relative proximity with every other segment. Part of the similarity between two segments may be due to a number of shared feature specifications. But since features at the phonetic level are generally assumed to be multivalued (rather than binary) the relative similarity of two segments also depends on scalar differences between their specifications along any given feature dimension.[1]

Matrix representation.-- The internal structure of phonetic representations assumed by the feature-segment systems of traditional phonetics may be conveniently represented as a matrix:

> ...a phonetic representation has the form of a two-dimensional matrix in which the rows stand for particular phonetic features; the columns stand for the consecutive segments of the utterance generated; and the entries in the matrix determine the status of each feature.
>
> (Chomsky and Halle 1968:5)

Affinity structures and systems of physical interpretations.

The two aspects of feature systems.-- The specification of the relationships of partial identity and relative proximity that are assumed to hold among a set of phonetic segments will be referred to as the *affinity structure* of that set. While the above relationships are to be determined from "feature

specifications", we will distinguish between an affinity structure of a feature system and the feature system itself. The affinity structure is an abstract set of relationships, part of the formal structure of phonetic representations. To constitute a feature system, an affinity structure must be provided with a set of external conditions, or rules of *physical interpretation*, which relate it to a set of physical parameters.

Two (or more) distinct feature systems may share the same affinity structure. To illustrate this, we will consider an example that is relevant to later discussions. According to traditional phonetic analysis, the vowels [I], [ɛ], and [æ] are assumed to differ (primarily) with respect to their values along with respect to a single feature. Assume that the values 3, 2 and 1 are given as their respective specifications on a feature arbitrarily labelled "height". Traditional feature systems suppose that the affinity structure is related to a physical parameter based on the vertical position of the tongue in the mouth.

However, other physical interpretations of the same set of relationships are possible. Indeed, it has been proposed by Ladefoged (1971) that "height" relationships are realized as differences in the frequency values of the first formants of vowels such as those in the present example, and NOT as tongue positions.

It would be possible to supply details of each of these systems of physical interpretation so that they would make identical claims about the relative proximity relationships among this set of vowels. In such a case, the two different feature systems would share the same affinity structure.

Empirical evidence for traditional affinity structures.-- The distinction between a concrete, physically interpreted feature system and its affinity structure is a logically important one. Considerable evidence exists to support certain aspects of the affinity structures of traditional vowel descriptions. However, in the next chapter it is shown that there also exists considerable evidence AGAINST the traditional system of physical interpretations.

There are at least two types of evidence that may bear on the affinity structure of a set of phones while bypassing the question of the system of physical interpretations. The first of these is the patterning of groups of phones with respect to sound laws and in phonological patterns.

The phonetic relationships of classes of sounds have been considered relevant to diachronic change since at least the time of the neogrammarians. In synchronic phonology, American structuralist practice had recourse to a criterion of "phonetic similarity" for the grouping of allophones into (taxonomic) phonemes. Prague school theory (Trubetzkoy 1969) emphasized the oppositions of phonological elements along a limited set of phonetically defined dimensions. Generative phonology has put forth the relevance of phonetic feature structure to the phonological systems of language as perhaps its most fundamental principle (Halle 1964).

In these cases, it is not the physical properties of segments that is of primary importance, but rather the relationships of partial identity and relative proximity of sets of segments.

A second type of evidence that potentially bears on affinity structure (but not on systems of interpretation) are psychological experiments dealing with subjects' judgments of the similarity of pairs of phonetic items. These will be discussed in a separate section below.

It seems possible that affinity structures can be tested by phonological data and psychological "scaling" experiments. Systems of physical interpretation can only be tested by analysis and synthesis of actual physical parameters in relation to an affinity structure; that is, by experimental phonetics.

The relationship between physical parameters and phonetic features: the speech code

Physical scales versus grammar-mediated constructs.-- Perhaps the most common hypothesis about the nature of the relationship between phonetic features and physical parameters is that there exists a simple one-to-one correspondence between values on single physical dimensions and feature specifications. Thus Chomsky and Halle state, "The phonetic features can be characterized as PHYSICAL SCALES, describing independently controllable aspects of the speech event (1968:297)."

While phonetic representations are to supposedly consist of matrices of such features, the simple definition implicit in the above statement apparently does the listener little good since these authors also maintain:

> We might suppose ... that a correct description of perceptual processes would be something like this. The hearer makes use of certain clues and certain expectations to determine the syntactic structure and semantic content of an utterance. He uses the phonological principles he

> controls to determine a phonetic shape. The hypothesis will then be accepted if it is not too radically at variance with the acoustic material, where the range of permitted discrepancy may vary widely with conditions and many individual factors (1968:24).

Apparently, on this explanation, though phonetic representations are matrices of universal features, they can only be derived by the listener through language-specific components of the grammar.

Whether or not such a curiously inaccessable universality contains any logical contradictions,[2] such a model appears to be unable to account for several readily observable abilities of naive speaker-hearers.

Phonetic sensitivity and the problem of interference.-- Perceptual experiments of various kinds have obtained reliable responses to speech and speech-like stimuli with no syntactic or semantic structure whatsoever: nonsense syllables are reliably perceived. Similarly, new proper names can apparently be learned without extreme difficulty. Such facts argue against a strong version of a grammar-mediated construct account of listeners' perceptions. However, there is evidence that language-specific differences exist in the perceptions of naive listeners' catagorization of other languages (cf., e.g., Lotz, Abramson, Gerstman, Ingemann and Nemser 1960).

On the other hand, there are also indications that naive speakers have access to relatively detailed phonetic representations that cannot be supplied by higher level components of the grammar. We agree with Ladefoged that relatively narrow phonetic specifications are part of "...what makes an Englishman sound like an Englishman even when we are paying no attention to what he is saying (1971b:275)." The frequently observed experience of the detection of foreign accents and dialects by untrained listeners seems to depend in some measure on sensitivity to quite subtle phonetic distinctions.

Ladefoged (1971b) also cites an example of an area where the detection of a subtle phonetic difference is actually a requirement for the correct assignment of semantic content to a sentence pair: *I'm going to get my lamb prepared* versus *I'm going to get my lamp repaired*. Other "juncture" phenomena considered by researchers in the era of taxonomic phonemics would appear to depend on equally subtle phonetic differences.

Universality and the speech code.-- The position to be adopted here is that phonetic representations are universal precisely because they depend solely on language-independent relationships between physical parameters and phonetic features. We further

accept Mattingly and Liberman's conclusion that "...the interconversion of phone and sound is an integral part of language and of its underlying physiology" (1968:111). The system that specifies this interconversion is referred to by Mattingly and Liberman as "the speech code."[3] This concept of "interconversion" implies that phonetic representations are (ultimately) recoverable from the acoustic signal. This will be referred to as the principle of the *acoustic recoverability* of phonetic information. It is meant to imply nothing more than that the acoustic signal contains all the information necessary for the recovery of phonetic information by human listeners. It is not meant to imply the "primacy" of acoustic parameters over articulatory. The question whether an articulatory representation is a necessary intermediate step in the recovery of phonetic information from acoustic signals will be returned to below.

Phonetic features as functions of physical parameters

Relative parameter values.-- Though the definition of phonetic features as "physical scales" seems to be compatible with many traditional statements, there are at least some instances in which more complex relationships between features and parameters seem to be implied. The traditional account of tone is a case in point. For example, Pike's (1943) comments on the specification of tone levels in tone languages imply that this feature is determined by the relative values of pitch parameters distributed over time rather than being a simple isomorphic mapping of instantaneous, absolute pitch.[4] He remarks:

> ...tones are "high" or "low" relative to each other rather than to an absolute pitch; changing the key, i.e. all the pitches, does not change tone; a single, level, isolated tone is not subject to classification -- it must be put next to others to see if it is relatively low or high and so on (1943:27-28).

The uniqueneness of phonetic representations.-- It seems to be a tacit assumption of traditional phonetics that every utterance has one and only one phonetic transcription. This may be called the principle of the uniqueness of phonetic representations. This principle of uniqueness together with the principles of universality and acoustic recoverability implies that phonetic representations are UNIVERSAL FUNCTIONS OF ACOUSTIC PARAMETERS.

Such a notion avoids the extreme limitations of the definition of phonetic functions as physical scales. At the same time, it constitutes a stronger (more constrained) theory of the relationships between phonetic representations and physical parameters than theories that hold that phonetic representations are available as grammar-mediated constructs.

Matrix formulation of phonetic functions.-- Phonetic representations have been defined above as matrices of features, where the columns represent segments and the rows represent feature specifications. We may also regard the acoustic signal as a matrix, where the rows represent relevant (to the phonetic theory in question) acoustic parameters and the columns represent samples of those parameters taken at relevant time intervals. Formally, the notion that phonetic representations are functions of acoustic parameters may be stated as follows: there exists a universal (species-specific but not language-specific) function T such that for every matrix P of relevant acoustic parameters the function T specifies one and only one matrix F of phonetic features.

The specification of the function T in the above statement, together with statements about what constitutes "relevant" acoustic parameters, actually corresponds to an entire phonetic theory. We will assume that T is analyzable into a number of (largely independent) simpler functions. Some of these may perform tasks of segmentation and others may map certain subsets of parameters to specific features.

The specification of features as functions of sets of physical parameters allows for features such as tone to depend on relative rather than absolute values. The formulation of the functions in terms of matrices of parameters and matrices of features relieves us from the requirement of a one-to-one association between single instants of time and single segments.

Returning to the details of the remarks on tones above, we note that the information required to arrive at a tone specification may be spread out over a relatively long-term context. Such a situation would constitute a minor problem for the strict definition of phonetic features as functions of physical parameters. When the long-term context is sufficiently detailed, the "key" (to use Pike's musical analogy) of the tones can be determined and it is presumably possible to establish a unique tone specification for a syllable on a given pitch (in a given position in an intonation pattern). However, in certain situations when the range of pitches has not been sufficiently sampled, there should be ambiguous cases. The definition of a feature as a function requires that we be able to arrive at a unique specification of the phonetic feature for every relevant set of physical parameters. The existence of ambiguous, non-unique cases would seem to vitiate the proposed definition.

However, this is not so if we consider that the ambiguity is temporary rather than intrinsic. We will assume that the long-term information is part of the information required by the tone extracting function. When the information in the signal is

(temporarily) insufficiently specified with respect to these long-term parameters, we will assume that it is supplied by "hypothesis". Such a hypothesis may be supplied by expectations about syntactic form of semantic content, or perhaps by a guess as to the identity of the speaker.

This proposal is quite different from that of Chomsky and Halle mentioned earlier in this chapter. According to the latter, semantic and syntactic analysis is REQUIRED for the extraction of the phonetic structure of an utterance. The present proposal involves only the occasional specification, by hypothesis, of a limited class of long-term parameters which will ordinarily be forthcoming from the input signal. A "wrong" hypothesis about the underspecified parameters can be corrected as more information becomes available.

Articulatory features.-- The acceptance of the principle of the acoustic recoverability of phonetic representations does not entail the automatic acceptance of acoustic properties as "basic" to phonetic features. It is logically possible that the "key" to the speech code is a set of articulatory parameters. The mapping from sound waves to features may be accomplished through an intermediate association of articulatory properties to acoustic parameters. Indeed, many arguments in the last two decades for the "motor theory of speech perception" (Liberman et al. 1962) have assumed this to be the case.

Such a process is compatible with assumptions made above if phonetic features are thought of as being derived by two layers of functions: the first set mapping from sound to articulation, and the second set from articulation to phonetic features.

However, such a two-layered set of functions can usually be formulated as a single set of functions of the acoustic parameters alone. In order to motivate an intermediate stage, it seems necessary to argue: 1) that the articulatory parameters are simply related to the phonetic features; and 2) that the relationship between the articulatory parameters and the acoustic output is fully determined by universal physiological constraints (together with the laws of acoustics).

An example of features as functions.-- A consideration of the results of an experiment presented in the classic study of Cooper, Delattre, Liberman, Borst and Gerstman (1952) will serve to clarify the issues raised above. In the experiment to be considered, a series of synthetic CV syllables consisting of a set of two formant vocalic stimuli preceded by a noise burst were presented to subjects. The vowels ranged over the sounds [i, e, ɛ, a, ɔ, o, u] and the center frequency of the burst was varied over a range of 360 to 4320 hertz. The subjects were asked to

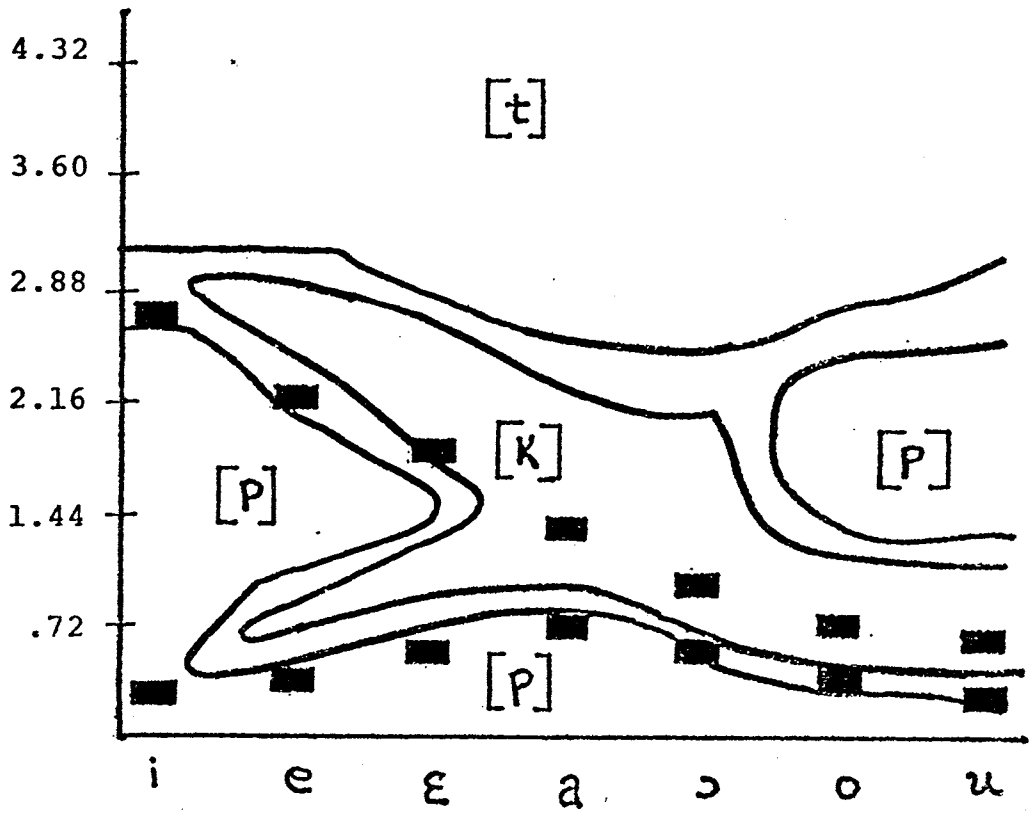


Figure 1.1.1. Categorization of stop consonants as a function of a noise burst and formant frequencies of following vowel. After Cooper et al. 1951. Vertical axis: center frequency of noise burst in kilohertz.

identify the initial sounds of the synthetic syllables as [p], [t] or [k].

A graph of the results of this experiment has been reproduced here as Figure 1.1.1. This graph characterizes the specification of the traditional place of articulation feature as a *function*[5] of the relationship between the center frequency of the noise burst and the formant frequencies of the vocalic portions of the stimuli. While no isomorphic association between the absolute value of the frequency of the burst and the value of the place feature can be made, the relationship between such a burst and the formant frequencies of the following vowel generally specifies a unique value for the feature. Cooper et al. characterize the basic pattern of this function as follows:

> We see that high frequency bursts were heard as *t* for all vowels. Bursts at lower frequencies were heard as *k* when they were on a level with, or slightly above, the second formant of the vowel; otherwise they were heard as *p* (1952:273-274).

The complexity of the relationship between the presumed phonetic feature and the acoustics should not obscure the fact that the feature is a function of the acoustic parameters. But this complexity has been used quite plausibly in arguments for the articulatory motivation of sound-to-phone relationships. One of the earliest arguments in this vein is presented by Cooper and his associates in the following passage:

> The results of the *PTK*-burst experiment ... provide some extreme cases which suggest that the perceived similarities and differences between speech sounds may correspond more closely to similarities and differences in the ARTICULATORY domain than to those in the ACOUSTIC domain; that is to say, the relationship between perception and articulation may be simpler than the relation between perception and the acoustic stimulus. (1952:605)

Evidence that there are articulatory gestures that regularly occur in the production of stop consonants of different place specifications is readily available even through unaided visual observation. Even so, for the notion of the primacy of articulation to attain full force, it would seem necessary to demonstrate that the complex acoustic pattern is accounted for solely by universal consequences of a set of invariant articulatory parameters.

Whether or not a detailed model of "acoustic diversity from articulatory unity" can be maintained for consonants is a question beyond the scope of this current study. However, it appears that there is substantial evidence against such a model in the case of

vowels. Rather, it would appear that a set of relatively constant acoustic relationships are preserved in the vowel systems of different speakers in spite of considerable variation in the articulatory patterns used to produce them.

On the acoustic recoverability of articulatory information

The above discussion has served to illustrate that the principle of acoustic recoverability does not necessarily imply that phonetic features are "more directly related" to acoustic rather than to articulatory events. What it does imply is that if features are to be viewed ultimately as functions of articulatory parameters, these articulatory parameters must themselves be derivable as functions of properties of the acoustic signal. It is therefore important to consider the relationship of various aspects of articulation to acoustic signals.

It appears that there are substantial difficulties in establishing a FUNCTIONAL relationship in the direction of sound waves to articulatory properties. What this would require is that the relevant articulatory information is uniquely recoverable from acoustic events. While it would appear that there is a unique mapping from articulation to acoustic output, the mapping from acoustic signals to articulation would seem to involve non-unique mappings at several levels. The extent of the uniqueness problem depends on the nature of the articulatory parameters that are held by a given theory to be most directly related to phonetic features.

There are three levels of articulation that will be considered here. Proceding from more central to more peripheral, they are: 1) motor commands; 2) configuration of articulators; and 3) area functions.

Traditional feature specifications such as place and manner of articulation may be regarded as hypothetical claims for the specification (or "physical interpretation") of phonetic features in terms of the configurations of certain articulatory structures. An articulatory configuration is the result of the interaction of active muscle forces with various passive constraints imposed by the "hardware" of the vocal tract. The muscles are controlled by motor commands. For a fixed set of passive constraints (i.e., for one individual's vocal apparatus), given an initial configuration. it seems reasonable to suppose that the subsequent configuration will be uniquely determined by a given set of motor commands.

However, it is not clear that the relationship is uniquely determined in the opposite direction. Because of the complexity of musculature of the tongue, different degrees of antagonism, synergy, etc., it seems reasonable to speculate that the same configuration of articulators might be produced using different

muscle commands. Though this author is aware of no evidence directly bearing on this question, it is at least questionable in principle whether motor command information is recoverable even from a full history of articulatory configurations within a single individual. When variation in passive constraints is added to account for differences in individuals' vocal tracts, the hypothetical problem is compounded.

Acoustically, at any moment in time, the only important aspect of an articulatory configuration is the area function it specifies. An area function may be represented by a graph displaying distance from the glottis on the x-axis; the cross-sectional area of the vocal tract is shown on the y-axis for each point on the x-axis from the glottis to the lip opening. While each configuration of the articulators will determine a unique area function, the same area function could result from more than one configuration.

This possibility may be illustrated by a hypothetical example. For certain vowel articulations the vocal tract is said to approximate a uniform tube (Fant 1960). In the case of a perfectly uniform tube, the graph of the area function would appear as a straight line parallel to the x-axis. But the length of such a tube could be varied at either end: by protruding or retracting the lips; or, by raising or lowering the larynx. Thus two gestures, while taking place at opposite ends of the vocal tract would have exactly the same kind of effect on the area function.

Finally, while (assuming a constant source of excitation) the acoustic output is uniquely determined by a given area function, the same output can evidently be produced by distinct area functions. Thus according to Atal (1974:1), "the acoustic information cannot, in general, be mapped in a one-to-one manner into an area function without imposing additional constraints on the articulatory mechanism". It seems to be a crucial theoretical question for any articulatory-oriented feature system whether any such constraints exist in actual speech situations.

To summarize, it appears that given an initial configuration and a constant glottal source, acoustic output is completely determined by motor commands. The mapping in the other direction appears to involve, in principle, a multi-layered indeterminacy: the same acoustic output can result from any of several area functions; each of the area functions are possible outcomes of several distinct articulatory configurations; these in turn may be the result of different sets of motor commands. As we move from more peripheral manifestations to more central ones (from

area functions to motor commands) the problem of unique specification appears to be compounded.

Whatever level of articulatory representation is taken as basic by a particular model, it ultimately falls on such a model to specify those constraints that allow the relevant articulatory parameters to be uniquely derived from acoustic signals. Perhaps dynamic constraints, say, on the changes from configuration to configuration will ultimately be shown to meet the acoustic recoverability condition. However, evidence presented later in this work makes it appear unlikely that this is the case for vowels. In the next chapter it is shown that very similar acoustic events are produced by quite different patterns of articulation for different speakers.

Traditional feature specifications for vowels

For the remainder of this work we will be concerned with the range of phonetic variation generally referred to as vowel "quality" or "color". In particular, we will concentrate on the phonetic distinctions covered by the traditional features "advancement", "height" and "rounding". These correspond to the quality features of the system of the International Phonetic Association (1949), hereafter abbreviated IPA. The feature "tenseness" which seems generally less well imbedded in traditional systems will also receive somewhat less attention.

The terms "advancement", "height", and "rounding" will be used in a sense that is strictly neutral with respect to physical interpretation. Traditionally, these terms have an association with aspects of tongue and lip position, but they will not be so used here unless an articulatory prefix is explicitly included. Thus, "tongue height" is to be distinguished from "height" in that the former implies a particular system of physical interpretation.

Unprefixed terms are mnemonic labels for dimensions of a hypothetical affinity structure. These dimensions may be delineated by reference to the vowel segments (expressed as alphabetic symbols) they are assumed to differentiate.

The feature advancement is taken to distinguish such vowel pairs as [i] vs. [ɨ] or [a] vs. [ɑ]. The terms "back" and "front" are used to characterize opposite extremes of the advancement scale. Advancement is traditionally given a physical interpretation related to the horizontal position of the tongue.

The feature height is taken to be the primary phonetic dimension of phonetic distinction between a series of vowels

such as [ɪ] - [ɛ] - [æ] or [ɯ] - [ʌ] - [ɑ]. The terms "high" and "low", or equivalently, in the IPA terminology, "close" and "open", are used to characterize extreme opposite specifications. Height is traditionally interpreted as "tongue-height", a physical parameter relating to the vertical position of the tongue in the mouth.

The feature rounding is taken to distinguish such vowel pairs as [i] - [y] or [ɯ] - [u]. Extreme values are often characterized simply by the terms "round" and "unround". However, IPA terminology provides for values ranging from "close rounding" through "open rounding" and "neutral" to "spread". Rounding is traditionally interpreted as a physical parameter related to various aspects of lip position.

The feature tenseness is sometimes taken to be part of the distinction between vowel types such as [i] and [ɪ] in German. Extreme values are characterized as being "tense" and "lax". The physical interpretation given traditionally is one of generally greater (muscular) tension for "tense" vowels. The feature is not recognized as a dimension in the IPA system.

Another feature that will occasionally be referred to is one that includes the range of phonetic contrast covered by both the advancement and rounding features. This feature will be called "clarity". The vowels [i] - [y] - [ɯ] - [u] are sometimes treated as though they varied along a single dimension, with [i] described as maximally "clear" or "acute" and [u] as maximally "dark" or "grave".[6] Such a clarity dimension is a prominent feature of a tradition of vowel description from which the familiar vowel "triangle" diagrams have apparently been derived. Statements in the traditional literature are generally vague as to its physical interpretation, though they generally seem to indicate that it is an auditory rather than an articulatory distinction. (See, for example, Trubetzkoy 1969, pp. 97-104.) The general topology of vowel diagrams in which a clarity dimension is present indicate that it may be related to the frequency of the second formant.

There is a difference in the affinity structures of analyses that propose a three-dimensional height by advancement by rounding system as opposed to those which propose a two-dimensional height by clarity system. Empirical evidence for the independence of advancement and rounding features will be considered below.

Acoustic parameters

The frequencies of the first two formants are usually considered to be the primary acoustic determinants of vowel quality (Fant 1959, Peterson 1961). It has also been proposed that the third

and sometimes higher formants have a role to play in the formation of an "effective second formant", called by Fant (1959), F2'. Values of F2' have been estimated in an experiment reported by Carlson, Granstrom and Fant (1970) in which subjects were asked to match a two formant pattern to four synthetic reference vowels. While the matched second formant was generally placed fairly close to the second formant of the reference stimulus, in some cases, particularly for vowels in the high front area, it was placed considerably higher.

The important point to be noted here is that a series of two formant patterns corresponding approximately to two formant patterns in natural speech appears to be SUFFICIENT to produce all the vowel distinctions covered by the traditional advancement, height and rounding features. This fact was first indicated by the work of Delattre, Liberman and Cooper (1951). It is corroborated by the perceptual experiments reported by Carlson et al. (1970) for Swedish listeners, since the Swedish vowel system contains a wide variety of feature distinctions which may be conveyed satisfactorily by two formant patterns.

A plot of values for estimates of formant frequencies appropriate for a set of IPA cardinal vowels provided by Delattre et al. (1951) is reproduced here as Figure 1.1.2.

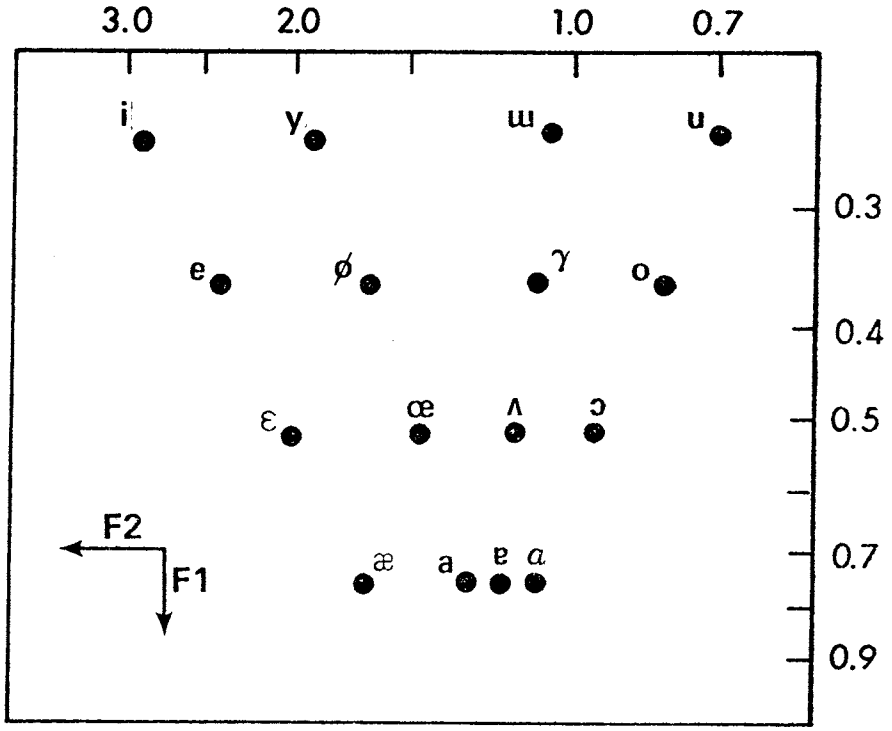


Figure 1.1.2. Acoustic plot of major vowel categories. After Delattre et al. 1951. Axes in kilohertz.

Scaling of similarity judgments

A number of techniques for extracting "underlying perceptual dimensions" that have been developed in recent years have been applied to subjects' judgments of the relative similarity of speech sounds. These procedures, consisting of various types of factor analyses and multidimensional scaling, will be referred to collectively as scaling procedures. While a detailed discussion of such methods is not possible here, some of the results that have been reported for studies of vowel sounds deserve brief consideration. Important aspects of the affinity structure posited for vowel quality are corroborated in the relationships of vowels along the "perceptual dimensions" extracted by these methods.

In all of the seven vowel studies he reviews, Singh (1974) finds that the first two hypothetical perceptual dimensions extracted correspond roughly to (rotations of) height and advancement features. In only one of these studies, that of Terbeek and Harshman (1971) is there a suggestion of a rounding feature. This occurs in the analysis of data from German speakers; two other groups of subjects, Thai and English speakers, did not show this dimension.

While many of the studies reviewed by Singh show evidence for more than two perceptual dimensions, only two are consistently interpretable across languages and despite variation in the nature of the stimulation (natural and synthetic speech, isolated vowels and vowels in word contexts). While it would appear that one of these dimensions corresponds to height, the second dimension which Singh interprets as advancement may correspond better to a clarity feature. While a plot of height by advancement would be expected to show vowel pairs such as [i] and [y] or [e] and [o] with roughly the same values, this does not generally appear to be the case in the actual analyses. The studies of Hanson (1967) for Swedish and of Pols, van der Kamp and Plomp (1969) for Dutch show relationships among such vowels on the first two perceptual dimensions to be somewhat more similar to what would be expected on an F1 by F2 plot, or on a height by clarity plot. That is, the relationships are generally in line with the diagram of Delattre et al. (1951) shown in Figure 1.1.2.

Phonetician's transcriptions and phonetic representations

The transcription experiment of Ladefoged (1960). -- Ideally, a phonetician's transcription should contain all and only the universal phonetic information of an utterance. Actual phonetician's transcriptions may be thought of as attempts to "measure"

phonetic information. Though there is some evidence that transcriptions do not under all circumstances reflect purely universal phonetic information, in certain controlled circumstances, quite consistent "measurements" have been obtained.

The concern that field phoneticians' transcriptions may be influenced by the phonetic inventories of their native dialects has been put forward by Ringgard (1965). Witting (1961) presents evidence that transcriptions reflect phonotactic patterns (combinatory constraints) of the transcribers' dialects. Witting's subjects were a group of Swedish students in their first year of phonetic training.

But evidence exists that experienced listeners trained in the same tradition arrive at quite consistent phonetic descriptions of vowel sounds despite differences in language backgrounds.

Ladefoged (1960) reports an experiment in which 18 linguists were presented with high fidelity tape recordings of 10 Gaelic words produced by a native speaker. The linguists were asked to plot the vowels on a "cardinal vowel chart." The coordinates of such a chart correspond generally to height and advancement judgments and may be thought to constitute "analogue transcriptions" of those features. Subjects were also asked to record rounding judgments. None of the listeners was familiar with the dialect of the speaker.

Fifteen of the 18 phoneticians were "trained in the British tradition of phonetics." Though there was considerable diversity in the language and dialect background of this group, their transcriptions showed considerably less diversity than those of the other three subjects. We will limit the rest of the discussion to the judgments of the British-tradition phoneticians.

For seven of the 10 vowels, variation in the advancement by height plane was quite small. While differences between the vowels were large, Ladefoged found that the mean area containing at least 14 of 15 judgments for each vowel was only about two percent of the area of the vowel diagram.

The other three vowels showed somewhat more variability. Interestingly, for the two most variable of these, Ladefoged found a significant correlation in the degree of rounding and the relative advancement in the transcriptions. Ladefoged takes this to indicate that "... the degree of lip rounding is not always considered an independent variable"(1960:395).

The experiment of Laver (1965). -- Laver (1965) presents the results of the phonetic judgments of 10 synthetic, steady-state vowels by five "British-tradition" phoneticians. The data recording technique was similar to that in the experiment of

Ladefoged (1960) described above. Each of the phoneticians transcribed each of the 10 vowels a total of 40 times in the course of five days. The experiment included a number of "camouflage" vowels so that the listeners were unaware that they had categorized the same synthetic token more than once. There was some degree of variability in the repeated recordings of the same token by the same listener. Laver remarks: "... the variations of location seemed, on the whole to show no overall pattern-shift, but rather a random movement about the average location"(1965: 113). He further notes that this variability presents some difficulty for claims of extreme accuracy in individual acts of transcription.

On the other hand, Laver's study shows that the average position of the advancement by height plots over the forty tokens show excellent agreement across the five subjects. Hurford (1969), using Laver's data, calculates the mean variability in this experiment for the 10 vowels to be on the order of only .85 percent of the area of the vowel chart.

Transcribability.-- The experiments of Ladefoged (1960) and Laver (1965) indicate that phoneticians' judgments are derivable from acoustic properties of speech and speech-like events. While more experiments would be desirable, these results indicate that human language meets what might be termed a "transcribability" condition.

If, as assumed in this work, phonetic representations are functions of physical parameters, the fact that language is transcribable is not surprising. If, on the other hand, phonetic representations are supposed to be recoverable from the acoustic signal only after a complex grammar-mediated matching process, the degree of transcribability manifested in these experiments seems inexplicable.

Advancement and rounding versus clarity

Though the experiments discussed above indicate a high degree of transcribability for height and advancement features, both experiments noted a relatively higher degree of variability in rounding judgments. This variability assumes a greater importance in light of several other phenomena.

First, as noted earlier, while quality distinctions are assumed to vary in three articulatory dimensions, all such distinctions appear to be recoverable from a TWO-dimensional acoustic space, F2 by F1. Secondly, a perceptual dimension corresponding to the rounding distinction does not generally emerge in scaling

analyses of listeners' similarity judgments. Thirdly, as discussed in the next section (2.2), for as long as 80 years before the three-dimensional articulatory classification, vowel quality differences were displayed in a two-dimensional plane, corresponding roughly to an F1 by F2 plot. In the two-dimensional system, a single dimension, clarity, was used to distinguish vowels of the same modern height classification that differ in either "tongue advancement" or "lip-rounding" or both.

Summary of section 1.1

In this section we have proposed that phonetic representations be considered as matrices of features that are derivable ultimately as functions of matrices of acoustic parameters. There are experimental indications that human language meets a transcribability condition, at least in the case of vowels. The lawful, though sometimes complex, relationships between acoustic parameters and listeners' categorization of synthetic speech stimuli are also consistent with the notion of a transcribability condition.

We have also indicated that certain aspects of the affinity structure of traditional feature analyses for vowels are supported by statistical analyses of naive listeners' judgments of the similarity of speech sounds. In the next section, we will investigate the history of phonetic descriptions for vowels concentrating particular attention on their affinity structures.

1.2 Affinity structures and Feature Systems in the History of Impressionistic Phonetics

Introduction

What follows is a historical outline of several of the impressionistic phonetic schemes that have been proposed for the description of vowels. There are two basic points to be made in this outline. The first is one that has been noted by Ladefoged (1967): in spite of the widespread use of the three-dimensional articulatory feature system for the description of vowel quality, there has been considerable variation in accounts of the articulatory properties of vowels in the history of phonetics. The second point is that there exists another tradition for the representation of vowel quality which has been assimilated in the IPA system which has remained relatively stable since the end of the 18th century. This is the "triangle tradition" which appears NOT to have been organized in articulatory terms.

Pre-18th century descriptions

J. A. Kemp in his introductory comments to a reproduction and annotated translation of the 17th century work of John Wallis (Wallis 1972) outlines the feature systems for vowels of 11 pre-18th century phoneticians. Kemp notes: "Articulatory descriptions of vowels in the 16th and 17th centuries vary considerably in the categories [i.e., features] they start from and the total number of vowels they allow for"(Wallis 1972:43). While many of the systems of classification he outlines include a basic "place" by "aperture" classification, they are frequently modified by other features including tongue shape and larynx height.

Many of these descriptions do not appear to include "tongue height" parameters for what is now described as the back rounded series of vowels, but rather these vowels are distinguished by lip parameters alone. Ladefoged (1967) discusses a few of these systems. Excellent reveiws of early vowel descriptions are provided by Michaelis (1881) and Vietor (1898).

Perhaps some of the variation noted by Kemp can be ascribed to the fact that the group he discusses includes phoneticians of different times and different language backgrounds. However, even in the works of four nearly contemporary 17th century British phoneticians, Wilkins, Wallis, Holder and Cooper, there seems to be considerable variety in the articulatory descriptions of what must be essentially the same set of vowels.

While a detailed comparison of the differences among these authors will not be attempted here, it is suggested that the comments of Holder, written in 1699, may provide an explanation for such divergences:

> The Articulations, that is the Motions and Postures of the Organs in framing the Vowels, are more difficultly discerned, than those of the consonants; because in the Consonants, the Appulse is more manifest to the sense of Touching but in the Vowels it is so hard to discern the Figures made by the Motions of the Tongue, (inclining onely toward the Palat, and not touching it) especially about the more inward Bosse or Convex of it that it is rendered no less difficult to define the articulations of the vowels, and he that can describe them accurately, *erit mihi magnus Appolo* (Holder 1967:82-83).

This passage is followed immediately with an astute methodological observation concerning introspective judgments: "Onely he who shall adventure, has this advantage, that it is easier to affirm than to disprove"(p. 83).

C. F. Hellwag and the "German vowel triangle"

The "triangle tradition" in phonetics is generally agreed (Michaelis 1888, Lazicius 1961, Zwirner and Zwirner 1970) to have started with the 1781 Tubingen dissertation of C. F. Hellwag (Hellwag 1781), entitled *De Formatione Loquelae*.

Hellwag is partly interested in anatomical descriptions of speech production. However, his introduction of the triangle seems to be motivated primarily by concern for some type of auditory, rather than articulatory, relationships. The following passage is offered as this author's translation of Hellwag's section 57 in which the famous figure is introduced.

> The first of the vowels, the base of the rest, is *a*, to be placed in the center of a scale. From it two scales rise, terminated in their extreme degrees by *i* and *u*: intermediate terminals are set between these extreme degrees and between corresponding ones below. The relationships of the intermediate degrees and terminals to the base can be represented by the following figure: [Hellwag's diagram is reproduced here in Figure 1.2.1]
> ...
> The vowel *o* holds a middle place between *u* and *a*; and å between *o* and *a*. Similarly, *e* lies between *i* and *a*; and ä between *e* and *a*. There is a transition from *u* to *i* through ü; and one from *o* to *e* through ö. A terminal point could be provided through which the transition from å to ä is made. Among these degrees designated by writing, it is possible to interpolate countless others, which the different peoples pronounce in their languages and a varieties of languages. Is it not the case that all vowels and diphthongs ever to come forth from the human tongue can thus be determined quasi-mathematically according to degrees?

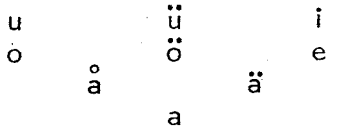


Figure 1.2.1. Hellwag's (1781) vowel diagram.

That some type of auditory judgment serves as the primary basis for the relationships described seems to be indicated by the first sentence in Hellwag's section 58 which follows immediately the above: "not only does listening commend these regular rows of degrees, but a careful consideration of changes in the mouth confirms them."

Hellwag then continues with some description of the articulatory configurations of the vowels, concentrating on the vowels *i*, *a* and *u*. The vowel *a* is taken to be, in effect, a maximally open vowel with "quiescent" lips and tongue and a maximally lowered jaw. The vowel *u* has lip rounding, minimal jaw opening and the "root of the tongue" is said to be maximally elevated towards the back. The vowel *i* is produced with minimal jaw opening, quiescent lips, and minimal space between the palate and the tongue. He describes *ä* and *e* as being articulatorily between *a* and *i*, while *å* and *o* are articulatorily between *a* and *u*. As in modern descriptions, *ü* and *ö* are described as rounded versions of *i* and *e*.

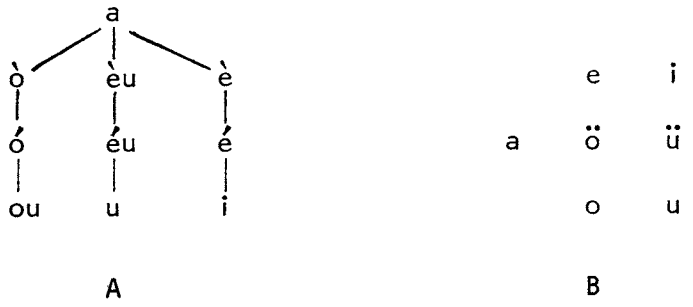
The betweenness relations indicated in the articulatory descriptions would seem to correspond in a straightforward manner to the vertical axis of his articulatory description triangle. As in later descriptions, this axis can be related to a putative aperture measure. The main difficulty in the reconciliation of his articulatory descriptions with the triangle diagram lies in the interpretation of the horizontal axis. It appears to correspond to what we have called a clarity dimension.

Imposing Cartesian coordinates, *a* would be assigned the same coefficient on the abscissa as would *ü* and *ö*. Since Hellwag recognizes the independence of rounding and tongue position in the articulatory sphere, there seems to be no articulatory explanation for this partial identity in the affinity structure implicit in his diagram. However, after his discussion of the articulations of vowels, he offers an ordering of the sounds of several of the vowels on a scale of most grave to most acute as follows: *u o å a ä e i*. Although he does not interpret *u* and *o* on this scale, he has provided for the other seven vowels a SINGLE dimension, defined in auditory terms, which corresponds exactly on an ordinal scale to the relationships among those vowels along the X-axis of his triangle diagram.

We will assume that Hellwag's *ü*, *ö*, *ä* and *å* represent vowels of roughly [y], [ø], [æ] and [ɒ] qualities respectively. His other vowels, *i*, *e*, *a*, *o* and *u* will be assumed to represent qualities near those represented by similar symbols in modern IPA transcription. If this is so, then the ordinal relationships among the six vowels in Hellwag's acute-grave series correspond to relationships in F2 in the vowel diagram of Delattre et al. (1951) presented in Figure 1.1.2. Furthermore, the vowels [y] and [ø] stand in roughly the same relationships in Hellwag's triangle along the clarity axis as they do along the F2 dimension of the formant plot.

Whether or not Hellwag had intended an exact correspondence between his articulatory descriptions and the triangle, it is clear that as the tradition continued, a non-articulatory interpretation prevailed. This is evidenced by the various rotations and reflections that the figure apparently underwent in the first few decades after Hellwag. The acoustician Chladni (1809) presented the diagram with French orthographic symbols reproduced here in Figure 1.2.2(A). This may be seen to be a slightly squared-off, rotated and reflected version of Hellwag's diagram.

Dubois-Reymond (1812) provides the diagram reproduced here as Figure 1.2.2(B). We learn from Vietor (1898) that a nearly identical figure (with the same orientation) appears in an unpublished manuscript, dated 1780, found in Hellwag's literary estate. This fact lends credence to the position that the figure was not originally intended by Hellwag to represent articulatory relationships.



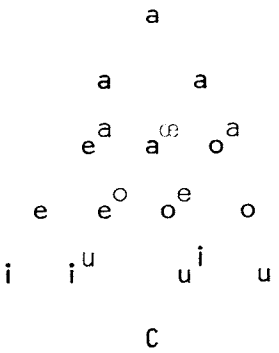


Figure 1.2.2. Nineteenth century vowel diagrams.

A. Chladni (1809)
B. Du Bois-Reymond (1876)
C. Brücke (1867)

Elaborations of the *a*-topped vowel "pyramid" seem to have been the most widely used on much of the continent throughout the 19th century (Lazicius 1961:63). The diagram of Brücke (1876) is presented in Figure 1.2.2(C). The use of such a pyramid extended well into the 20th century, as it is the form used by Trubetzkoy (1969).

In sum, while some minor modifications in vowel inventory and slight differences in the shape of the figure appear in the course of the triangle tradition, the basic relationships among the vowels remain relatively stable. The following relationships appear widely in the vowel diagrams of this tradition:

1) [i], [a] and [u] appear at the vertices of a roughly triangular figure.

2) The figure is symmetrical about the perpendicular bisector of the [i]-[u] line.

3) If Cartesian axes are imposed, [i] and [u] always share a coefficient on one of the coordinates. In effect, [i] and [u] "share a feature specification."

4) The vowels [i], [a] and [u] all have different values on the other Cartesian axis, with [i] and [u] showing maximal separation of any two vowels along that axis and [a] showing an intermediate value between the two.

5) Vowels of different modern rounding specifications may share the same coefficients along the latter axis.

These relationships characterize some important properties of what will be referred to as a *Hellwagian figure*.

Nineteenth century British phoneticians

A. M. Bell.-- In the earlier part of the nineteenth century, British phonetics seems to have been primarily concerned with problems of transcription and shorthand (Albright 1958). The first real activity in what might reasonably be called feature theory seems to have begun with Alexander Melville Bell.

The system of vowel description first expounded by Bell in *Visible Speech* (Bell 1867) is generally believed to be the primary source of modern articulatory theory. It is there that the terms "high, low, back, front, round" and "unround" first appear. However, a consideration of some of the details of Bell's system of physical interpretation makes it clear that there has been considerably more change in articulatory description from Bell's time to the present than is indicated by the similarity of feature names.

The system used by Bell before the appearance of *Visible Speech* bears a resemblance to several of the systems of the 17th century British phoneticians, though it is identical to none.[7] In *Visible Speech*, Bell describes this system:

> ...so recently as 1862, when a new edition of the "Principles of Speech" was called for, the author had not advanced beyond his original triple scale of vowels consisting of the three classes
>
> *Lingual* *Labio-lingual* *Labial*
>
> the first series starting with the close *ee*, the third with the close *oo* and the intermediate with the German *u*, and each series terminating in the most open vowel *ah* (1867:14-15).

This system appears to be basically compatible with the affinity structure of the Hellwag triangle. However, Bell goes on to describe his dissatisfaction with the way this system dealt with certain vowels of English and French in the [œ] - [ɜ] area. Contemplation of these problems led to the "revolutionary" re-analysis of the entire system of vowel description. He continues:

> The revisal of the "Principles of Speech" had reopened the whole question of elementary relations and the experimental classifications which followed, resulted in the identification of a new category of vowels, -- a series moulded simultaneously by the back and front surfaces of the tongue . . . .
>
> It was evident that there were three classes of purely lingual vowels, moulded respectively by the back, the front, and by "mixed" back and front surfaces of the tongue; and that each element in this triple scale was the basis of another vowel, in forming which a definite labial quality was simply added (1867:16).

While the definition of front and back vowels seems congruent with modern treatments, the definition of the lingual configuration of "mixed" vowels is quite different from that of their modern successors, "central" vowels. Bell later replaced the category "mixed" with the category "top", formed by a middle surface of the tongue (Bell 1897).

Though in the above passage lip rounding appears roughly similar to modern descriptions, in a later section of *Visible Speech* his definition of rounded vowels is quite different:

> All the varieties of ... vowels hitherto explained, result from the shape and size of the cavity of the mouth as affected by the Tongue, while the lips remain spread so as not to influence the sound. The same lingual positions yield another series of vowels when the voice-channel is "rounded" and the aperture of the lips contracted. The mechanical cause of "round" quality commences in the super-glottal passage, and extends through the whole mouth-tube, by lateral compression of the buccal cavities and the reduction of the labial aperture. The last cause -- lip-modification -- being the visible cause of "round" quality is assumed as representative of the effect (1867:76).

Bell also includes what is essentially a "pharyngeal feature", orthogonal to both tongue position and rounding:

> ...it was found that the cardinal degrees were amply sufficient for all practical purposes, in connection with another distinction which now revealed itself: the distinction between Primary Vowels or those most allied to the consonants, and the Wide Vowels or those in forming which the pharynx or gutteral passage is fully expanded (1867:71).

This feature appears to bear only partial correspondence to the "tense-lax" distinction in terms of the vowel pairs it was intended to separate.

Thus Bell's *Visible Speech* system retained only the "aperture" dimension of his earlier system unchanged. Three new features, horizontal tongue position, a pharyngeal modifier, and a complex rounding distinction, replace the earlier three way classification.

Henry Sweet.-- None of the new features proposed by Bell were to go unmodified in the system of his most influential pupil, Henry Sweet. Sweet's sometimes radical departures from Bell's system are not as obvious as they might be, since Sweet seems to have made a deliberate effort to modify his teacher's terminology as little as possible.

Sweet completely redefined Bell's primary-wide distinction, renaming it "narrow-wide." Perhaps owing to his (acoustically inaccurate) conviction that "... alterations in the shape of the pharynx have only a sound-coloring, not a sound-modifying effect (1910:463), "Sweet provided an oral tongue-based description to supplant Bell's pharyngeal distinction. According to Sweet the "narrow" vowels were produced with a kind of bunching of the tongue. This bunching was always accompanied by a somewhat

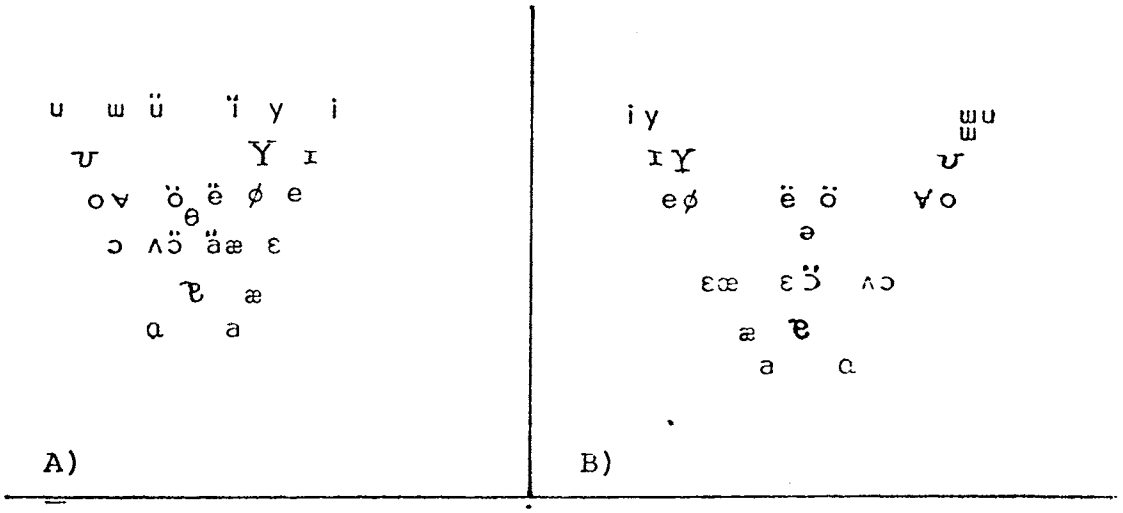
greater height or narrowing of the passage of the mouth.

Sweet (1971:79) added another dimension, which we will refer to as "slope", to the inventory of possible lingual distinctions. Bell's three horizontal values, "front-back-mixed" are preserved, though the two-bulge description of the mixed category is explicitly rejected. Each of these positions may occur in connection with, in effect, a marked and unmarked tongue slope. Back and front vowels are normally associated with the sloped shapes, while mixed vowels are normally flat. But "marked" combinations of slope and position may also occur. In fact, Sweet doubles the number of cardinal tongue specifications for non-labialized vowels by combining marked and unmarked slopes with the three horizontal positions.

Sweet apparently abandoned all non-labial aspects of rounding included in the system of his mentor, Bell. However, as late as 1906 he distinguished between two distinct types of rounding, "inner" and "outer" (Sweet 1971:62-63). Only the vertical tongue position specifications of Bell's system appear to have remained unmodified by Sweet.

The merger of the triangle tradition with articulatory specifications

Michaelis (1881) may have been the first to provide an explicit link between a Hellwagian figure and the "front-back" terminology of Bell and Sweet. Vietor (1898) was an early exponent of the linking of the "vowel triangle" with explicitly articulatory parameters. Otto Jespersen and Paul Passy, both very influential in the founding of the International Phonetic Association (Albright 1958) were generally favorable to the triangle diagram as a method of representation of vowel quality. Passy (1888) presents a diagram for French vowels in which rounded vowels are placed in parentheses next to the unrounded vowels with the same presumed tongue positions. However, a vowel diagram of Passy (1890), reproduced in Vietor (1898), does not pair the rounded and unrounded vowels in quite this way. Rather, the pattern is generally more like those of the "triangle tradition" diagrams in which rounded and unrounded counterparts are spread out along the "clarity" dimension. Passy's later diagram is virtually identical to that of the 1900 IPA figure (Albright 1958:55), reproduced here as Figure 1.2.4(A). This figure bears a striking resemblance to the acoustic diagram of Delattre et al. (1951; see Figure 1.1.2). The vowel diagram of the 1914 IPA (Albright 1958:56) again shows a pairing of rounded and unrounded vowels. See Figure 1.2.3(B).



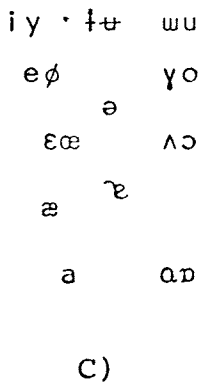


Figure 1.2.3. IPA vowel diagrams.

A) 1909; B) 1914; C) 1949.

The diagram of the 1949 *Principles* of the IPA is presented in Figure 1.2.3(C). The only change from the 1914 diagram (Figure 1.2.3(B)) is that the line of the back vowels is now perpendicular to the advancement axis. This modification appears to be the only direct contribution of empirical findings other than visual, tactile and/or proprioceptive observation to traditional vowel theory.

This modification is based on the 1917 X-ray photographs of Daniel Jones' production of the cardinal vowels. Four of these are reproduced as the frontispiece of later editions of Jones' *An Outline of English Phonetics* (e.g., Jones 1969). These modifications appear to have been incorporated in IPA diagrams in the late twenties, around the time of the appearance of Jones' (1928) monograph, which appeared as a special publication of the IPA. Curiously, the modifications of the vowel diagram appear to be considerably less radical than Jones' brief account of the X-ray results (Jones 1969:36-39) would suggest. Jones cites practical reasons for the simplification of his vowel diagrams. It is interesting to speculate whether the compromise figures have been chosen because a diagram based on the articulatory data would do violence to the affinity structure manifested by the older diagrams.

Summary of section 1.2

It can safely be said that all phonetic systems which have received any widespread use among linguists have been based on impressionistic rather than instrumental evidence. Aside from the occasional use of palatograms (which provide at best only very incomplete records of articulation) and the modicum of radiographic evidence in the case of Jones, all the articulatory systems discussed above are supported only by unaided visual observation and judgments of proprioceptive and tactile sensation. Changes in feature systems in so far as these have reached the general linguistic public have been based on the re-evaluation of subjective evidence available since the beginning and NOT on technological breakthroughs. Since there is no reason to believe that there should be any change in descriptions of proprioceptive and tactile feedback, the articulatory specifications of vowels can to some extent be questioned on the basis of historical evidence alone.

By contrast, the relative stability of the relationships represented in the diagrams of the "triangle tradition" seems noteworthy. The resemblance of the affinity structure implied by these diagrams to formant plots suggests that acoustic

relationships may more reliably characterize vowel systems than articulatory specifications. In the next chapter, empirical evidence is shown to lead to the same conclusions.

NOTES TO CHAPTER ONE

[1]Thus, if three segments [a], [b] and [c] have feature specifications on feature "X" of 1, 2 and 3 respectively but are otherwise identical, segments [a] and [b] are more similar than are segments [b] and [c].

[2]Difficulties with Chomsky and Halle's account of the feature extraction process have been pointed out by Ladefoged (1971a) and Catford (1974).

[3]Mattingly and Liberman emphasize the possible role of articulation as the "key" to the speech code. Such an emphasis is not LOGICALLY necessary to the basic notion of a regular and specific relationship between a complex set of acoustic events and phonetic representations. Whether it is empirically necessary is discussed briefly below.

[4]Whether this formulation is an accurate account of native listeners' perception is questionable. Abramson (1962) shows that the absolute fundamental frequencies of Thai speakers' production of syllables of various lexical tones is generally compatible with a relative pitch hypothesis. However, perceptual experiments by the same author (Abramson 1962, 1972) indicate that high identification rates are possible for words spoken in isolation. The following account of tone is intended merely as an example of the fact that relational hypotheses of feature specification are not entirely alien to traditional phonetics.

[5]Such a function involves the distribution over time of acoustic information necessary to specify the value of the place of articulation feature. This case differs from the hypothetical case of tone in several important ways. Perhaps the most important is that it involves relatively short-term relationships that are always present with stop consonants. The long-term relationships presumed to be important in tone specification are not always available to a listener.

[6]The terms "acute" and "grave" in this connection do not correspond exactly to the same terms as used in the feature system of Jakobson, Fant and Halle (1952) which are widely known among linguists. Therefore, the terms "clear" and "dark" will be used throughout.

[7]The system described by Bell seems to bear closest resemblance to that of Wilkins (1668), though Wallis' (1972) work is

generally thought to have had a strong influence. See the discussion on this point by Ladefoged (1967:64-67). Ladefoged also cites there the system used by Wheatstone (1837) which seems to be the same as Bell's early system quoted below.

CHAPTER II

VOWEL QUALITY FEATURES AND ARTICULATORY PARAMETERS

2.1 The Problem of Articulatory Invariance for Vowels

Introduction

In the words of K. Harris:

> The motor theory of speech perception is a statement that we will find a simpler relationship between a string of phonemes that a listener perceives and the articulation of the speaker, than between the acoustic signal the speaker generates and perception (1974:2281).

There are, as discussed below, two different theories of the organization of speech PRODUCTION that could underlie such a theory of perception. Such theories of production and perception taken together constitute models for the articulatory SPECIFICATION of speech. Such models claim, in effect that phonetic representations (which are neutral with respect to production and perception) are directly related to articulatory parameters. Models of articulatory specification are in fact models of articulatory invariance: they imply that phonetic features are associated with invariant aspects of the articulatory process.

Models of the articulatory specification of speech may be opposed to models of acoustic specification. Such models would claim that, although the speech signal is a result of an articulatory process, the organization of that process cannot be understood without reference to its acoustic consequences. Speech sounds are implemented articulatorily, but they are specified acoustically.

Many of the arguments put forth by advocates of models of the articulatory specification of speech have been arguments against the notion of acoustic invariance of phonetically equivalent events. In the discussion below, some of these arguments as they have been applied to vowels are surveyed. It is argued that some of the cases of severe variation in acoustic parameters cited as instances of lack of acoustic invariance may in fact be associated with phonetic variation. To the extent that variation in the acoustic signal is associated with variation in the phonetic message, there is no invariance problem at the PHONETIC

level. Furthermore, from the evidence reviewed below, it appears that to the extent a phonetic invariance problem exists for acoustic parameters, it appears to present equally severe difficulties for all models of articulatory specification thus far proposed.

Most of the arguments for lack of acoustic invariance have centered about acoustic variation within the speech of a single speaker. The problem of articulatory invariance is apparently at least as severe in these cases.

Evidence from the literature supplemented with new experimental data makes it appear that there is in fact a more severe problem for articulatory invariance when the case of cross-speaker variation is considered. In the acoustic data, the correspondence between physical parameters and phonetic events is more straightforward. The traditional advancement and height features are more clearly related to acoustic parameters of vowels than they are to the articulatory tongue positions they are ordinarily taken to represent.

Motor and configurational target theories of speech organization.

Models of articulatory invariance.-- MacNeilage (1970) notes that many writers "... regard the actual serial ordering of speech production primarily as the output of sequences of phoneme 'commands' according to higher order rules of the structure of language"(1970:183).[1] He further observes that the primary difficulty facing such a model is the problem of within-phone variation. His concise delineation of the most widespread approach to the resolution of this problem is contained in the following passage:

> These authors consider that the main problem that such a view must cope with is the well-known fact that acoustic correlates, and therefore by inference, vocal tract configurations, are known to exhibit enormous variability, due particularly to variations in phonological context, speaking rate and stress....
>
> With remarkable unamimity, the authors of phoneme-based models consider this lack of correspondence between the phoneme and its peripheral correlates to be the result of three factors: (a) the mechanical constraints inherent in the peripheral vocal structures; (b) limitations in the response capabilities of the neuromuscular system; and (c) overlapping in time of effects of successive phoneme commands. This peripheral variability is thus not taken to invalidate the possibility of a discrete invariant phonemic input in their models (1970:183).

MacNeilage notes further that there have been two major views advanced with regard to the nature of the articulatory invariance that underlies the time-smeared peripheral events. The first of these is the invariant motor command model; and the second, the invariant configurational target model of speech production.

Motor commands.-- The invariant motor command model is associated with the motor theory of speech production as advanced by Liberman, Cooper, Harris, and MacNeilage (1962). This theory held that "... EMG [electromyographic] correlates of the phoneme will prove to be invariant in some significant sense (Liberman, Harris, MacNeilage and Studdert-Kennedy 1967: 84, cited by MacNeilage 1970)." But MacNeilage contends that electromyographic research has actually indicated substantial variability for a given phone in different phonetic contexts. In fact, he concludes that "... the main result of the attempt to demonstrate invariance at the EMG level has been not to find such invariance but to demonstrate the ubiquity of variability"(1970:84). He further argues:

> ... the more basic problem in speech production is not the one considered essential to most theorists: namely, why articulators do not always reach the same position for a given phoneme. It is, How do articulators always come as close to reaching the same position as they do? One of the main conclusions of this paper is that the essence of the speech production process is not an inefficient response to invariant central signals, but an elegantly controlled variability of response to the demand for a relatively constant end (1970:184).

Targets.-- The invariant configurational target models of speech production have assumed that instructions to attain certain positions rather than motor commands are the underlying invariant properties of speech sounds. Since invariant positions are not achieved, as indicated by the variable acoustic consequences which result, target models require some additional theoretical apparatus to account for natural speech data. Perhaps the most frequently cited account of failure of target achievement is that of Lindblom (1963).

Reduction and stress effects

Passive coarticulation model.-- In studying the acoustic consequences of vowel reduction in unstressed syllables of a Swedish speaker, Lindblom concludes that the phenomena can be accounted for in terms of an exponential function of target

formant frequencies of consonants and vowels and of segment duration. He also found that the same model fit reasonably well to the data of the same speaker producing stressed CVC syllables at different speaking rates. Lindblom proposes that these effects could be accounted for by means of essentially passive modifications imposed on invariant target commands by the human vocal apparatus:

> A vowel target appears to represent some physiological invariance. The present data support the assumption that the control that the talker exercised over his speech organs in vowel articulation is associated with neural events in a one-to-one correspondence with linguistic categories. Let it be further assumed that an utterance is a sequence of such events that serve to trigger the appropriate articulatory activity. Articulators respond to control signals not in a stepwise fashion but smoothly and fairly slowly, owing to the intrinsic physiological constraints. Since the speed of articulatory movement is thus limited, the extent to which articulators reach their target positions depends on the relative timing of the excitation signals. If these signals are far apart in time, the response may become stationary at individual targets. If, on the other hand, instructions occur in close temporal succession, the system may be responding to several signals simultaneously and the result is coarticulation (1963:1778).

Lindblom further remarks that a speaker's "... strategy of encoding is clearly not intended for a listener who demands absolute acoustic invariance in the realization of phonemes, but it presupposes that the listener is able to correct for coarticulation effects" (1963:1780). While there is at least limited support that perceptual evaluation of vocalic stimuli is dependent on phonetic context and duration (Lindblom and Studdert-Kennedy 1967) in a manner generally consistent with the observed reduction effects of Lindblom (1963), there appears to be little physiological evidence to support the notion that acoustic undershoot effects for vowels are the result of passive coarticulatory constraints.

Distinct articulatory targets for reduced vowels.-- Houde (1967) presents strong evidence that vowels in unstressed syllables of English have actively different target positions from their stressed counterparts. In a cineradiographic investigation of the vowels [i], [ɑ] and [u] in nonsense words of the form [V_1 'b V_2 b V_1], he notes:

> The stressed vowels observed in this study were all at least 250 ms. in duration. Thus it may be assumed, on the basis of Lindblom's findings, that a vowel target position was reached during all stressed vowels. Unstressed vowels occurring in the middle syllable were generally too short to allow the attainment of a target position. However, the unstressed vowels in the initial and final positions of utterances were separated only by a short break in voicing (approximately 50 ms.) from the same unstressed vowels in the adjoining utterance. In this case the articulatory positions tended to be held through the voicing break. This sustained position (less than 1/2 mm change in 100 ms.) was identified as the unstressed vowel target position (1967:53).

In such a situation, one would expect that coarticulation effects are at a bare minimum. Yet Houde found the average position assumed by the tongue in the unstressed targets to be displaced consistently in a forward and upward position from that of their stressed counterparts.

This finding would appear to be partly consistent with traditional views of the nature of reduced vowels which generally hold that they are different in phonetic quality from their stressed counterparts. However, the direction of deviation in Houde's data is not that generally posited by traditional descriptions which hold that unstressed vowels (in English) assume a generally more centralized tongue position (cf. Lindblom 1963). We may seriously question whether stress-related reduction effects are in fact instances of "phone-preserving" variation. Here, the difficulties associated with the distinction between a "phone" and a "phoneme" assume a greater importance than elsewhere. Phonologically equivalent events are not necessarily phonetically equivalent events.

Context effects.-- The context-dependent formant variation in vowels noted by Stevens and House (1963) is apparently considered by them to be a special case of reduction effects. The overall effects on consonantal context show standard deviations of F1 and F2 on the order of five to seven percent in the means, except for F2 of [u] and [ʊ], which show somewhat larger deviations. Though this author is aware of no specific phonetic statements about American English that would account for the particularly large deviations in these cases, the possibility of extrinsic allophony should not be overlooked. Broad and Fertig (1970) in a large sample study (all possible CVC contexts) of the vowel [ɪ] also find systematic deviations in vowel formant

frequencies as a function of consonantal context. These effects are again rather small, showing standard deviations of about five percent in F1 and F2.

But three to five percent is the estimate of the magnitude of the just noticeable difference (JND) for formant frequencies of isolated vowels given by Flanagan (1955). Assuming the higher value and a normal distribution of the measurements, we would expect that about 95 percent of the values in the Broad and Fertig study fell within twice the JND of the mean values. It seems reasonable to ask whether a good deal of the variation noted for stressed vowels in different contexts is simply sub-threshold rather than "encoded" phone-preserving variation.

Further difficulties with passive coarticulation

Kuehn 1974.-- Speaking rate effects do not appear to offer any better prospects for articulatory invariance underlying acoustic variability than do stress effects. Evidence for changes in EMG activity as a function of speaking rate for two subjects is summarized by Gay as follows:

> ... Speaking rate effects cannot be attributed solely to articulatory sluggishness. The data of both Gay et al. (1974) and Gay and Ushijima (1974) show ... that vowels produced during fast speech are characterized by a DECREASE in the activity level of the muscle; in other words, undershoot is PROGRAMMED into the gesture (Gay 1974:265).

Kuehn (1974) also provides evidence that tends to run counter to passive undershoot models of speaking rate effects. As he notes, on such a model one might expect that articulatory velocities would remain constant. This is so because commands to articulators are assumed to be based on ideal target positions, whether or not they are actually attained. While Lindblom (1964) presents data from one subject that are generally in accord with the hypothesis, Kuehn's investigations indicate that such relationships cannot in general be maintained.

Kuehn concludes that at normal speaking rates "velocity of movement is contingent upon magnitude of displacement which depends on phonetic context within speakers and size of oral structures between speakers"(1974:2). He also reports that different subjects employ different strategies in change of speaking rate:

> In the present study, three subjects spoke more quickly primarily by increasing articulatory velocity. These

> subjects exhibited relatively little vowel undershoot. On the other hand, the other two subjects increased speaking rate by substantially reducing articulatory displacement (1974:101).

These observations make it appear unlikely that any fixed, universal processes underly the acoustic variability associated with rate differences. Furthermore, Kuehn's observations on the dependence of articulatory velocity on the size of the oral structures seems to place further obstacles in the path of undershoot models. He notes:

> A positive relationship was observed between articulatory velocity and tongue or jaw size: that is the larger the oral structure, the greater the speed and magnitude of movement (1974:1).

Note that this is NOT what we might expect from inertial coarticulation effects, if we assume that a larger structure is more massive.

Gay 1974.-- Gay (1974) introduces additional data that conflicts with undershoot models of coarticulation effects. His study combines cross-speaker comparisons for two kinds of factors which give rise to within-speaker variation: phonetic context and speaking rate. There are differential results for both effects in the two subjects and this study should serve as an indication of the need to test the generality of sources of variability on more than one subject.

In Gay's study, two subjects produced sentences containing nonsense trisyllables of the form [k V_1 'C V_2 p ə] where V_1 and V_2 included all combinations of the vowels [i], [ɑ] and [u] and C ranged over [p], [t] and [k]. The experimental corpus was repeated by each subject at a fast and a slow speaking rate. Articulatory results are presented in terms of the vertical displacement of a fleshpoint of the tongue, and the results of spectrographic analyses of the tokens are also presented.

Gay summarizes the effects of consonantal context on the slow speech data in the following passage:

> ... the vowel targets for both /i/ and /u/ are highly stable across changes in either the consonants or the vowel. The targets for /a/ are more variable, especially for one subject. Target position variability, when it does appear, is conditioned by both the consonant (left-

> to-right and right-to-left effects) and the first vowel (left-to-right effects). Right-to-left effects of the second vowel on the first vowel were virtually non-existent (1974:262).

In spite of the greater articulatory variability in the case of [ɑ], compared to [i] and [u], Gay notes "... acoustic variability for /a/ is no greater than for /i/ or /u/"(1974:265).

The nature of the articulatory patterns in the cineradiographic evidence in Gay's study shows speaking-rate effects which appear to be quite different from those noted by Lindblom (1964). While a general tendency for smaller articulatory displacements for all vowels in both subjects at the faster speaking rate is noted, (potentially an undershoot phenomenon), Gay notes: "The context effects that appeared at the slow speaking rate were generally absent at the fast speaking rate"(1974:262). Thus increased rate has actually resulted in a DECREASE in coarticulation. This result is clearly NOT what would be expected on the basis of inertial reduction effects.

Furthermore, and perhaps most remarkably, the observed undershoot effects do not result in "acoustic undershoot". Instead, Gay notes:

> For both subjects, an increase in speaking rate is accompanied by an increase in frequency levels of both the first and second formants .... The formant frequency measurements for both subjects show the same range of variation during fast speech as during slow speech ... [A]rticulatory undershoot during fast speech does not produce the same acoustic result as articulatory undershoot during destressed speech (1974:264).

Gay's overall conclusions on rate and context effects emphasize the acoustic output of articulatory gestures:

> Because variability is built into the production of a phone at a level higher than the peripheral speech mechanism, a vowel target cannot be internalized ... as an invariant event. Nonetheless, MacNeilage's (1970) three-dimensional coordinate system still seems to be the best basis for describing a vowel. However, such a specification would have to be expanded to include a spatial field, the boundaries of which are defined by the acoustic limits of the vowel (1974:265).

Summary of section 2.1

The main argument of this section may be stated as follows. There has been no successful attempt thus far to explain acoustic

within-phone variation in terms of observed physiologically invariant processes. Within-phone variation of the type discussed above appears to be no less problematic in articulatory than in acoustic terms. While this may indicate a basic weakness in the "target" notion of speech specification whether in acoustic or articulatory terms, it is also possible that the confusion of a number of distinct phenomena, some of which involve invariance only at the phonological level, has tended to exaggerate the magnitude of this problem at the PHONETIC level.

2.2 Configurational Targets for Vowels: Radiographic and Cineradiographic Evidence

Introduction

In the rest of this chapter, we will examine radiographic evidence that bears on the adequacy of configurational target specifications as phonetic features in conditions for which within-phone variation is not generally considered to be a serious problem; namely, in stressed syllables of slow speech for a single phonetic context. Traditional feature specifications are, in effect, hypotheses about invariant aspects of configurational targets. Since there appears to be independent evidence supporting aspects of the affinity structure of traditional theory, we will pay particular attention to the correspondence of traditional features to actual articulatory events. Some recent modifications of traditional systems will also be considered.

Russell's criticism

G. O. Russell (1928) presents the results of an extensive radiographic study of vowel production. By his own accounts, he had originally set out to verify the implications of the tongue arching triangle of Vietor (1898), an indirect forerunner of the IPA vowel diagram. But he was soon to become a strong critic of the traditional tongue-position feature system, rejecting both Vietor's system and that of the IPA on the basis of his empirical research. While some of Russell's objections have been partly answered in later research, others have been corroborated.

Russell's chief objections are summarized in a well-organized criticism of his work included in the radiographic study of Parmenter and Treviño (1932).[2] There are two main issues that these authors address. The first involves the question of articulatory target invariance within speakers and the second the affinity structure correspondence of observed tongue position to traditional phonetic specifications.

Repetition stability

On the question of target invariance, Russell claimed that there exists considerable variability in the production of the same vowel by the same speaker on different occasions. This may be referred to as the problem of the repetition stability of articulatory targets. Parmenter and Treviño note that most of the instances cited by Russell as examples of spontaneous token-to-token variability in fact were cases in which the subject had assumed noticeably different postures (as indicated by the slope of the back pharyngeal wall with respect to the hard palate) from one radiograph to the next.

Parmenter and Treviño undertake several radiographic experiments of their own in which the head position of the subject is elaborately controlled and systematically modified. They conclude that large postural changes can radically affect tongue positions. For a FIXED head position, they find that a single subject assumes almost exactly the same articulatory configuration on repetitions separated by as long as several months.

Additional evidence for the repetition stability of target positions is presented in the study of Carmody (1937). This work represents a report of radiographic experiments of R. T. Holbrook, "arranged and explained" by Carmody after Holbrook's death. Carmody comments on the fact that his late colleague had relied only on his "phonetic ear" and made no phonographic recordings of the subjects' productions. He remarks: "The films themselves prove the accuracy of his ear, for the many duplicates ... show identity of articulation even after an interval of several days or weeks"(1937:188). Carmody also quotes from a letter written by Holbrook which states: "the articulation for a given sound is almost always the same from film to film for a given individual no matter what the interval between films" (1937:232). [3]

Affinity structure correspondence

In addition to the question of target stability, Russell (1928) explicitly raises the issue of the correspondence of actually observed tongue position to those implied by phonetic tradition. While for the front series of vowels, Russell notes a tendency for there to be a narrowing of the palatal cavity and an expansion of the pharynx in the series [ɑ] to [i], he points out that "... it does not appear that we can postulate any universally regular progression from one of these vowels to the other, in the amount of arching against the hard palate in

front"(1928:278). He furthermore states that there are instances which clearly violate the rank order of openness relationships traditionally assumed in the front series of vowels:

> Our X-rays ... show that it is very common for the same subject to take a tongue position for the I (pip) which is more open than the e (pape).... As a matter of fact the I (pip) is sometimes more open than the ɛ (pep) and in some cases the I (pip) is actually more open than the æ (pap) all in the pronunciation of exactly the same individual (1928:333).

While Russell notes that the traditional order is sometimes found in the back vowels, he remarks: "Then so far as the back vowels are concerned, there is an even more shocking lack of conformity with our traditional designation of "open" and "closed" vowels ... there is actually more deviation than conformity"(1928:333).

A "defence" of traditional theory.-- Parmenter and Treviño suggest that most of the violations of traditional tongue position observed by Russell are due to the lack of control of head position in his experiments. They claim that in the subjects they studied, with head posiitons carefully controlled, "... there is a progression of the tongue as the vowels are produced in the traditional order"(1932:369).

In support of this position, Parmenter and Treviño present a radiographic study of ten American English vowels spoken by a single speaker. Their technique is to present composite tracings of each adjacent pair of vowels in the series [i, I, e, ɛ, æ, ɑ, ɔ, o, U, u], commmenting on the differences between them.

Interestingly, they seem to replicate some of Russell's findings: "... the highest point of the tongue is higher for [e] than for [I]"(p. 361). They also note that for [U] and [o], "the height of the tongue is about the same for the two vowels"(p. 361). They later conjecture that these difficulties may be due to the diphthongal nature of the vowels [e] and [o] in American English, and also with the "laxness" of [I] and [U].

With the exception of the two comparisons mentioned above they note:

> The comparison of the vowels in this series indicates a definite progression in the change of tongue position. For the front series i to æ, the front of the tongue which arches against the hard palate for i is lowered

> for each succeeding vowel, while the back of the tongue moves toward the back wall of the pharynx. In the back series u to ɑ, the back of the tongue rises and the opening between it and the velum becomes smaller (1932:362).

Notice here that the front and back of the tongue are to be used as references in determining height, rather than the traditional "high point". Clearly, from Parmenter and Treviño's own observations, these two measures do not always correspond. Furthermore, aside from the problems with [e] and [o], difficulties arise when one departs from the specific pairwise comparisons presented. Thus the superposition of the [U] and [ɑ] from this subject makes it appear that the two have essentially the same height in the back of the mouth, while [U] is actually lower in front.

Carmody's vowel figure.-- Carmody (1937) also attempts to defend traditional theory from the more radical aspects of Russell's criticism.[4] As a part of this defense, Carmody presents the vowel diagram reproduced here as Figure 2.2.1.

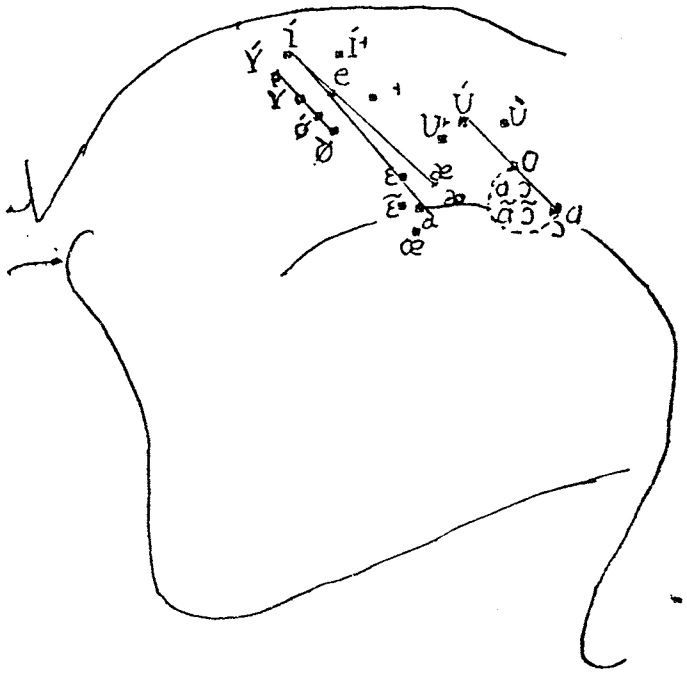


Figure 2.2.1. Carmody's (1937) vowel diagram.

He explains:

> ... It might be advisable once and for all to define what is meant by the "vowel triangle". Tradition has it (not unrightly) that the highest point of the tongue (its top bulge, never its tip) [but contrast this to Parmenter and Treviño's criteria above] offers the most convenient single identification of a vowel....
>
> To find a norm in our films, the "triangles" for each individual were drawn and then superimposed to form a single figure. Even with the inevitable differences in the size of jaw between individuals, it was found that most articulations fell on a single set of three lines, which form three sides of a quadrilateral (1937:231).

Note that this quadrilateral agrees with that of D. Jones (mentioned above) in the modification of the slope of the line of the back vowels. However, there are further modifications that are apparently unprecedented. The rounded front vowels appear distinctly forward of their unrounded counterparts.[5] Further, the lowest front vowel appears to be [œ]. There are additional qualifications noted by Carmody:

> [The low back vowels] ... depend mostly on lip position for their distinctive quality and so must be merged into a vague field which bounds their variations ... u varies little in height, but in some speakers it is formed farther forward than usual. English ʌ is too variable to locate without further material, since in our two tracings it falls once inside the quadrilateral, once directly behind o. English I is articulated by Mr. H. and Mr. Z. on the line, but by Mr. F. and Mr. L. in the same spot as ϕ^{+} (1937:231).

Cross-speaker variation.-- The individual variations mentioned in the above passage introduce perhaps the most profound problem for traditional phonetic specifications: that of the lack of CROSS-SPEAKER invariance. Further indications of such variability are given in Carmody's introductory remarks to a study of French vowels:

> Each speaker shows individual traits, often contradictory to those of other speakers; all articulations should be considered typical, and points of difference should be treated rather as unessential or variable characteristics than as departures from any fixed norm (1937:199).

But it is difficult to see just how the individual characteristics are to be distinguished from the essentials. Carmody's description of the back non-nasalized series of French vowels for four subjects appears to corroborate many of Russell's complaints:

> Change from ɑ to ɔ appears only in decreased jaw opening (15 mm. to 8 mm.) in D.'s speech, but it is not so consistent for R. Only very slight rounding characterizes C.'s ɔ, but B.'s more characteristic articulation [In what sense it is "more characteristic" is not clear, except that it corresponds better to traditional statements.] shows lip projection of from 3 to 5 mm., a 3-mm. rise in tongue, and a backward movement of from 3 to 4 mm. for ɔ.
>
> . . . . . .
>
> For ɔ and o, B. shows identical tongue position and jaw opening, but lips are more rounded for o, the lower being raised 9 mm.. For o, D. raises tongue 4 mm. toward velum, draws its tip back 12 mm., closes lips 5 mm., and extends larynx downward 5 mm. R. rounds lips, whether jaw opening is the same ... or closer for o..... [She] raises tongue 5 mm. toward velum and lowers larynx 4 mm. for o. C. shows no difference except a 6 mm. rise of vocal cords and a 4 mm. rise of lower lip....
>
> The change from o to u is noticeably consistent: all speakers raise tongue toward velum, move pharyngeal tongue bulge forward, and increase lip rounding. The most interesting of these changes is the great increase in the volume of the pharynx, its downward lengthening of from 5 to 8 mm. and the forward movement of the tongue in the pharynx 8 to 15 mm. (1937:203-207).

The nature of the acoustic differences in the series of vowels [ɑ, ɔ, o, u] is a simultaneous lowering of F1 and F2. The apparent trade-off in such various lip gestures as closure and protrusion with lowering, while diverse from the articulatory point of view, would appear to have a UNIFIED acoustic goal. Other things being equal, we would expect a simultaneous lowering of F1 and F2 for lip protrusion or larynx lowering (Stevens and House 1961, Fant 1960).[6]

Cineradiographic evidence for cross-speaker variability in vowel production

A possible objection to the relevance of earlier X-ray studies is that they are based on "unnaturally" sustained articulations because of the limitations of still radiography.[7]

Modern cineradiographic research, however, confirms the cross-speaker variability of articulatory parameters in vowel production.

The cineradiographic study of Ladefoged, DeClerk, Lindau and Papçun (1972) represents a major advance in the study of speaker-dependent differences in vowel production. Besides the empirical contribution, the authors focus attention on several critical theoretical issues. Their main conclusion, in line with earlier statements of Ladefoged (1967) that the traditional features of tongue height and advancement "... are more clearly correlated with acoustic rather than articulatory measures"(1972:74).

The authors present mid-saggital tracings of the five front vowels of six male speakers of what is adjudged by three phoneticians to be "the same variety" of American English. The frames chosen for tracing were determined on the basis of acoustic criteria from the spectrographic analysis of synchronized recordings on the following basis:

> In the case of the lax vowels in *hid*, *head* and *had*, a steady state part of the second formant was selected. For the tense vowels in *heed* and *hayed* which for some speakers were diphthongal throughout, a point shortly after the first consonant was selected (1972:56).

Data for the six subjects is discussed separately since "... there is a considerable degree of variation in the articulatory gestures used by the different subjects"(p. 64).

Tongue height and jaw opening.-- Included in this general variability is a replication of Russell's (1928) finding that the height relationships of the vowels [I] and [e] are sometimes contrary to those assumed by traditional theory. For four of the six subjects studied, the vowel in *hayed* is higher than the vowel in *hid*, on their measure of height.

Ladefoged and his colleagues also address certain recent modifications of traditional theory. Lindblom and Sundberg (1969, 1971) have suggested that jaw opening is a major factor in the articulation of vowels. However, from the data and discussion presented by Ladefoged and his colleagues, it appears that relative jaw opening is a very poor candidate for an invariant feature of vowels, since speakers vary widely in their use of this articulatory parameter. The rank order of the vowels along this dimension for each of the six subjects shows that it is even less satisfactory than tongue height in separating vowels across subjects.

Tongue shape.-- A second recent modification of traditional theory is that presented by Perkell (1971). Perkell presents cineradiographic evidence bearing on the suggestion that the feature "tense" appearing in Chomsky and Halle (1968) be replaced by a pair of pharyngeal features: advanced tongue root (ATR) and constricted pharynx (CPH).[8] Once again, there appears to be considerable individual variation in the pharyngeal configurations of different subjects. Concerning differences between the vowel pair [i, e], specified +ATR (by Perkell) and the pair [I, ɛ], specified -ATR, Ladefoged et al. (1972) note that three of their subjects show a wider pharyngeal region for the presumed +ATR pair while the other three do not. They remark:

> Subject 4 separates /i/ from /I/ by advancing the tongue root, but not /e/ from /ɛ/; subject 5 makes little use of the tongue root mechanism; and subject 1 uses it as a part of the mechanism for varying the tongue height for all vowels, both tense and lax (1972:72).

These authors also investigate the possibility of a general tongue bunching (whether achieved by tongue root advancement or not) as the basis of the assumed tense-lax phonetic distinction. For most subjects, the differences in shape among the vowels appear to be quite small and two of the subjects "... have essentially the same tongue shape in all the front vowels"(1972:72). Only three of the other subjects show a division between "tense" and "lax" vowels on the basis of tongue shape and for only one of these is the division described as "clear".[9]

An electromyographic study of the "tense-lax" distinction in the vowels of three speakers of American English reported by Raphael and Bell-Berti (1975) also shows considerable variability across speakers. While they find that two muscles investigated showed an increase in EMG activity for the tense vowels in all subjects, they point out that for the other ten muscles investigated "... no consistent tense-lax opposition is apparent"(1975:71). They continue:

> Further, each subject evidences at least one reversal of the hypothetical tense-lax difference for one of the muscles studied. Finally, even when the data reveals a tense-lax difference for a given vowel pair, that same subject frequently reveals either no consistent difference and/or a reversal of the hypothesized difference for another vowel pair (1975:71).

Summary of section 2.2

It would appear that neither the traditional features nor recent modifications of them stand up to empirical tests. The degree of inter-subject variability in these studies should indicate the inherent danger in basing a feature system on the data of a single subject. Indeed, empirical research since the thirties has produced evidence that weighs heavily against the notion of invariant articulatory specification in anything like that implied by traditional phonetic theory.

2.3 A Cineradiographic and Acoustic Study of 11 English Vowel Nuclei Spoken by Three Subjects

Introduction

The evidence reviewed so far in this chapter has indicated a substantial amount of cross-speaker variability in phonetically similar vowels. While acoustic parameters are known to vary substantially for speakers in different sex-age classes (see Chapters III and IV below), there appears to be only moderate variability for F1 and F2 measures among, say, adult male speakers of American English (Potter and Steinberg 1950). The experiment described below presents the results of a cineflourographic and acoustic analysis of a relatively wide range of vowel sounds from three adult male speakers. These analyses indicate that the acoustic parameters, F1 and F2, correspond better to the affinity structure relationships of traditional phonetics than do measurements of tongue and lip position. Furthermore, the cross-speaker variability of the articulatory parameters appears to be relatively severe when compared to that of the acoustic parameters.

Subjects and materials

The subjects were three adult males who are native speakers of slightly different varieties of American English. Though these dialect differences must contribute to the variability of the data, this author's impressions, supported by the transcription of an experienced phonetician (see below) indicate that the degree of phonetic variation among the speakers is relatively mild. Dialect differences will be noted wherever they appear to bear on the discussion.

Because we will occasionally be concerned with narrow phonetic description, in the rest of this chapter a distinction between narrow transcription, placed in square brackets, and a broad transcription, placed between slash marks, will be maintained.[10]

The vowels to be analyzed are /i, I, e, ɛ, æ, ɑ, ɔ, o, U, u, ʌ/, spoken in /b__b/ frames. These frames were in turn imbedded in the

sentence frame "It's a ____." The sentences were spoken in a pseudo-random order for subjects FSC and GNS and in the order the reverse of the above list for subject TMN.

Data recording

Lateral view X-ray films were recorded at a speed of 64 frames/second on 16 mm Kodak Plus-X film at the Eastman Dental Center, Rochester, N.Y. Three lead pellets (2.5 mm in diameter) were attached to the tongue of each subject with a surgical adhesive. The relative placement of the pellets of the subjects is indicated in Figure 2.3.1. The pellets will be denoted by their ordinal position, starting from the most posterior, P1, P2, P3. (Subjects FSC and GNS also had pellets attached near the tongue tip.) Pellets were also attached to the lips of each subject. The acoustic signal was recorded on magnetic tape using high quality audio equipment. Synchronization of the audio and cine data is discussed below.

Sound synchronization and frame selection

Although the audio synch-pulse generator at the Eastman Dental Center was not functioning on either of the dates of the X-ray runs, a selection of syllables bounded by bilabial stops enabled quite precise synchronization of the acoustic and cine data. Opening and closing of the lips was estimated at half frames when the change of a single frame resulted in differences between clearly open and clearly closed lip positions. Comparisons of the durations of the estimated time of opening and closure from the X-rays with the acoustic data indicated that errors of synchronization were not larger than one frame for the entire syllable.

Quantized and non-quantized broad-band spectrograms were made on a Voiceprint sound spectrograph. Most measurements were made solely on the basis of the standard (non-quantized) spectrograms, though the quantized versions were used where formant definition was poor.

Selection of frame.-- Frames chosen for analysis were initially located at points of minimal apparent movement as determined by eye from the films viewed at slow speeds. Rechecks indicated that in almost every case this system resulted in choosing the same frame for most vowels. The criterion initially employed was to find the earliest point of maximal lip opening. This usually corresponded to the point of minimal tongue motion as well. In a few cases where the acoustic record indicated a different point of minimal change, the articulatory material was

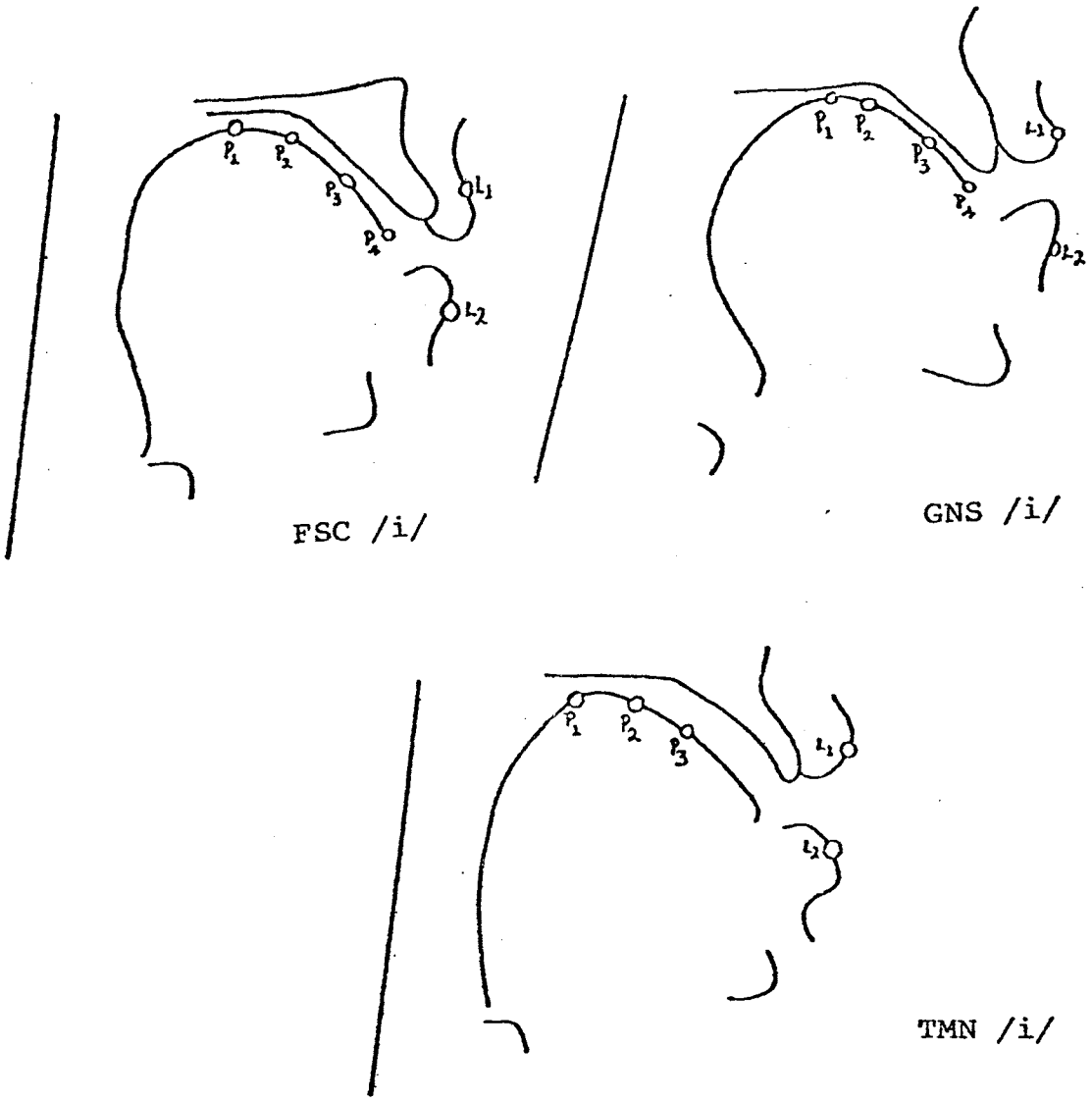


Figure 2.3.1. Relative pellet placement for three subjects.

re-examined. The discrepancies between acoustic and articulatory minimum velocities of change involved cases where tongue and lip motion were not entirely synchronized. The most difficult case was the segment /e/ for subject TMN. Maximal lip opening for this token did not occur until several frames after the point of minimal tongue motion. Since this vowel is phonetically diphthongal [eɹ], it is possible that part of the lip-opening gesture is actually associated with the off-glide, rather than being part of the lip-opening gesture to the main nucleus. Since the lip motion is relatively slow after the point of minimal tongue motion and fairly rapid before it, the point of minimal tongue motion was selected for tracing in this case.

Point of acoustic measurement.-- Points measured in the spectrograms were determined using information about opening and closing gestures of the lips in both the acoustic and cine data. This was accomplished by the following formula:

$$S_m = S_e + (S_i - S_e) \times (F_t - F_e) / (F_i - F_e)$$

S_m is the point to be measured in the spectrograms, S_e and S_i are the points of the spectrographic indications of initial /b/-explosion and final /b/-implosion respectively. F_t is the number of the frame selected for tracing; and F_e and F_i are the numbers of the frames for initial /b/-explosion and final /b/-implosion respectively.

The effect of this formula is best illustrated by an example. Suppose lip opening following initial /b/ began at frame 20 and ended (at the point of final /b/-implosion) at frame 35. The total duration of the opening is 15 frames. On the spectrogram, suppose we found a difference of 230 msec between the acoustic evidence of the opening and closing gestures (corresponding closely to the 234 msec predicted by 15 frames at 64 fps). If the frame chosen for tracing were 26, it would correspond to a point 6/15 or 40% of the total duration of the opening after the initial opening point in the cine data. The acoustic measurement would then be taken at 40% x 230 msec = 92 msec after the point selected as the time of initial /b/ explosion on the spectrogram.

Tracing procedure and equipment

The cine data was traced using a 16 mm Tage-Arnø film analyzer. Pieces of acetate sheet were taped to the viewing screen. Tracings were made on the acetate with a type of marking pen designed for use with overhead projectors. While the lines drawn were rather broad, they appeared to be adequate for the

level of detail discernable from the tracings.

While the quality of the films was generally good, the soft palate and the surface of the back wall of the pharynx were frequently not clear enough to allow reliable tracing from the single frames. When the films were viewed at slow speeds, however, these structures could be readily discerned. For all three subjects, the center line of the soft palate, except for its backmost section, appeared to assume a position that was very nearly an extension of the line of the hard palate.[11] The center line of the palate near the pharynx was harder to see.

Though the back wall of the pharynx exhibited some slight forward movement in all three subjects, particularly for the vowels /ɔ/ and /ɑ/, the variation appears to contribute little to the overall constriction of the pharynx in the mid-sagittal plane. An averaged back wall was constructed for each subject as a part of the "fixed structures" templates described below.

Templates were constructed for each subject from the superposition of about a dozen independent tracings of the hard palate, the outline of the upper teeth, the floor of the nasal cavity, and visible portions of the rear pharyngeal wall. Other prominent features of the skull were added where their contrast and shape made it appear that they would aid in the alignment process. The superposed tracings generally displayed excellent agreement. A single acetate sheet was then traced with dotted lines from the superposed tracings. Three or four arbitrary points beyond the areas of interest were added on the template as an aid in superposing the coordinate system described below for measurement. The acetate template was first aligned and taped to the screen. A blank acetate sheet was then taped over this for the actual tracing of a frame.

The maxillary coordinate system

A coordinate system for the recording of parametric measures was established for each subject on the following basis. On the composite "fixed structures" tracing, a line coincident with the relatively flat portions of the most posterior hard palate was extended rearward. This line was used as the X-axis of the coordinate system. The origin of the system was set at the point of intersection of the X-axis with a line fitted to the average back wall of the pharynx for the subject in question. See Figure 2.3.1.

It is not necessarily a trivial problem to construct a system in which measurements for one subject can be compared to those of another. Fortunately, the subjects involved in this experiment all have vocal tracts of nearly the same size, and palates of roughly the same shape. Figure 2.3.2 presents a graphic

justification for the use of the maxillary coordinate system. The X-axis and origins of the three subjects' systems are alligned. The hard structures and vocal tract configurations are traced for the vowel /i/. There is a divergence in the back wall of the pharynx of approximately seven degrees between subjects GNS and FSC.[12]

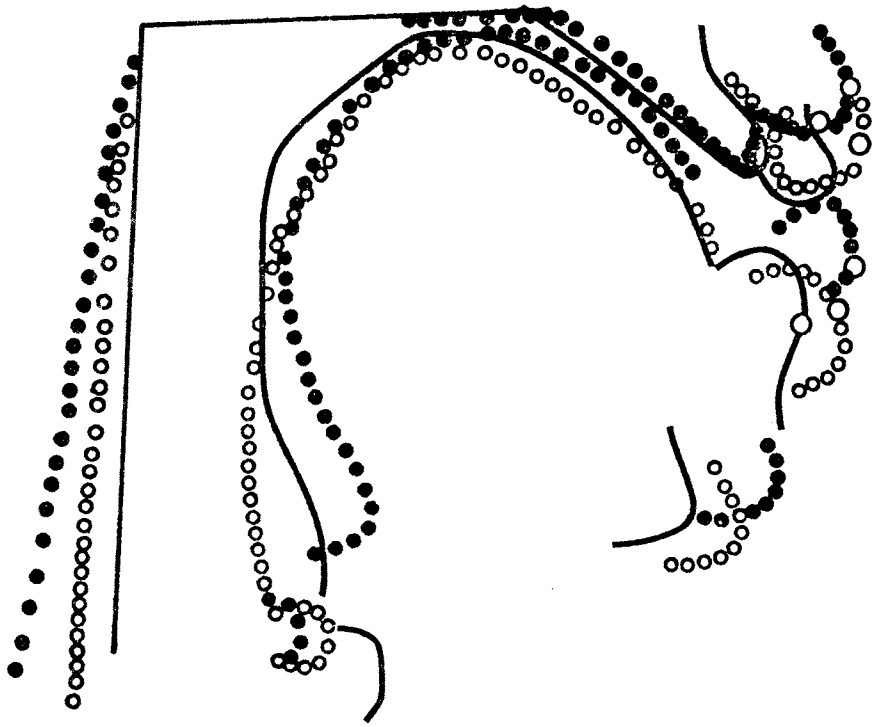


Figure 2.3.2. Superposed [i] for three subjects.

Parametric descriptions of tongue and lip positions

While some facts can be effectively illustrated by means of composite tracings, some kind of parametric reduction of information is desirable in order to summarize the general configuration of the articulators in the sense implied by traditional phonetics. Parametric measures used in this study are described below.

1) Flesh points as tracked by lead pellets.-- The chief difficulty with this measure is that the placement of pellets varied somewhat

from subject to subject. Furthermore, they were not always easy to locate in the single frames. The overall contour of the tongue was generally much easier to trace with confidence than the pellet position since they were continuous curves. The pellets presented special difficulties when they were near fillings in the subjects' teeth. However, the pellets could easily be followed when the films were viewed at slow speeds. When a pellet position could not be confidently determined from a single frame, its course was observed at slow speeds, with the hard structures of the tracing template aligned with the film. The apparent intersection of the pellet trajectory with the contour of the tongue as traced from the single frame was taken as the pellet position. An average position of the backmost three pellets was calculated for each vowel of each speaker and plots generally indicate that this average pellet position is a more stable measurement across subjects. It is not, however, a measure of pure position since differences in shape can affect the average as well as differences in the overall position of the tongue. Pellet positions for the lips were also traced. However, they did not seem to give a good indication of overall lip position.[13]

2) Highpoint of the tongue.-- Although this is the traditional measure of tongue position, tradition has not been very explicit as to its determination. The method adopted here is similar to that used by Ladefoged et al. (1972). The X and Y coordinates of the point on the tongue contour nearest to the X-axis of the maxillary coordinate system (see above) are recorded. A plot of the raw measurements is included in Figure 2.3.3. Since the raw plots appear to be substantially more variable than raw plots of other measures of tongue position, no further consideration of this measure is given.

3) Circle fits to the tongue and lower lip.-- The techniques of representation about to be described are modifications of parametric representations used in several articulatory synthesis schemes (cf. Mermelstein 1973). There are three measures that fall into this category, discussed in A, B and C below.

3A) Optimal circular bands fitted to the tongue contour. This technique represents an attempt at a measure of tongue shape. A series of concentric circular bands, corresponding to a width of about 3 mm each (in the original "life-size" scale) were drawn on an acetate sheet. Alternate bands were colored with a translucent pen to facilitate fitting. This sheet was placed over each tracing to find the circular band that contained the greatest length of the tongue contour with the point of maximum curvature of the tongue contour near the center of the fitted area. The size of the circle and the angle of the contour

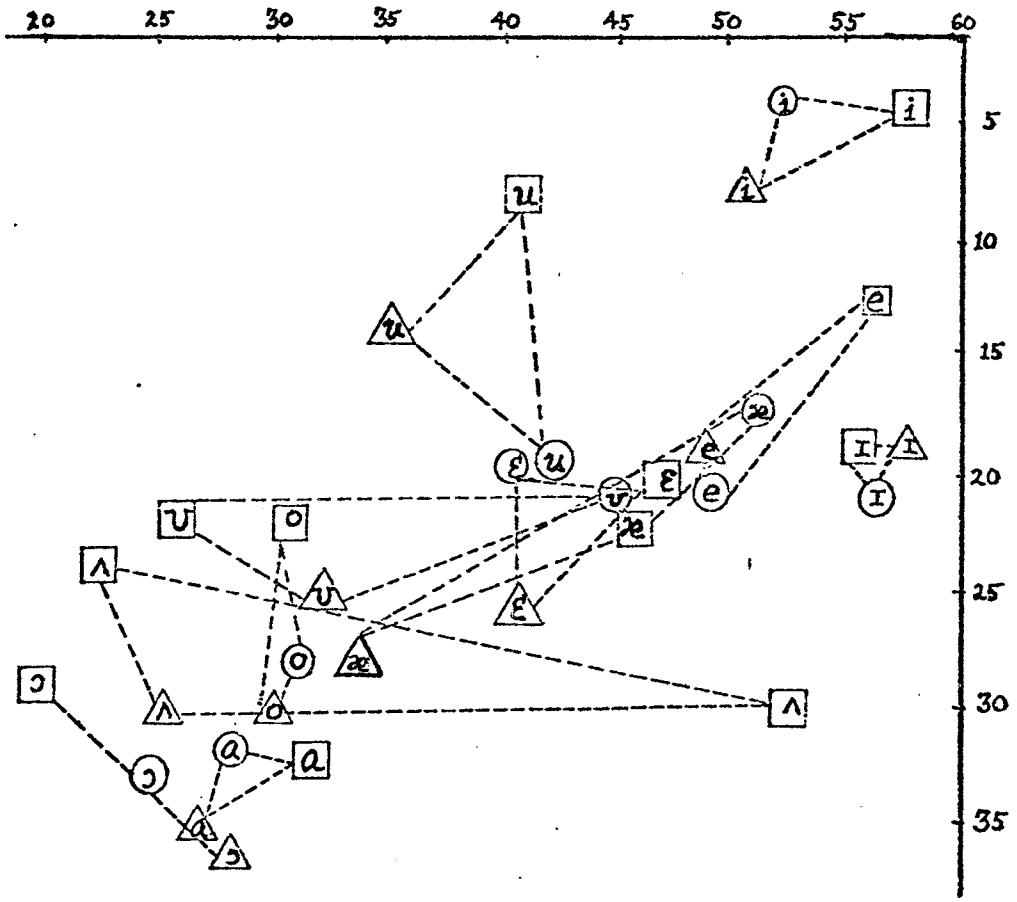


Figure 2.3.3. High point of the tongue. Scales in millimeters from origin of maxilary coordinate system.

(in 5 degree increments) contained within it were recorded. Other things being equal, a smaller circular band OR a larger contained angle indicates a generally "more bunched" tongue shape.

3B) Fixed circle fit to tongue contour. A circle of fixed radius somewhat larger than that of the optimal circular band for the three subjects was positioned so that its apparent fit to the tongue contour was maximized. Repeated measurements of the same tracings by such a procedure usually displayed variations of less than 1 mm. This method seems to give a better indication of the overall position of the tongue than any based on a small number of points. Since similar parametric representations of tongue position have been used successfully as major factors in the determination of area functions of vocal tracts in articulatory synthesis schemes, it would seem that this measure includes much of the information about tongue position that is most important in determining acoustic output. Because of the shapes of the tongue contours and of the hard structures of the vocal tract, the position of the center of the fixed tongue circle would also seem likely to correspond (except for constants on each axis) to the position of the "tongue pass" or the center of the region of maximal constriction in the vocal tract.[14] For an illustration of tongue circle fit, see Figure 2.3.4.

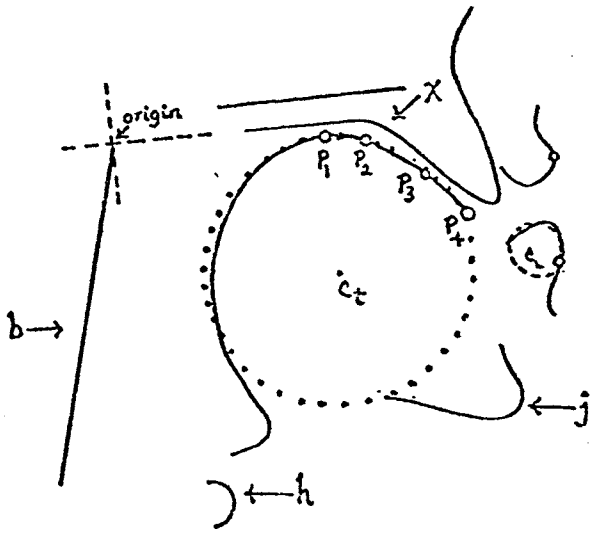


Figure 2.3.4. Illustration of elements of parametric description. C_t: center of fitted tongue circle; C_l: center of fitted lip circle. Other items of interest also shown; b: back wall of pharynx; x: hard palate; h: hyoid bone; j: mandible. "Origin" indicates the origin of the maxillary coordinate system.

3C) Fixed circle fit to the lips. A procedure similar to that described above was applied to the lips. This appears to be a more sensitive measure of the overall position of the lips than are the lip pellets since the lip in some instances appears to "rotate outwards" rather than protrude as a mass. The pellets, being positioned near the axis of rotation, show less displacement. Because of differences in the sizes of the lips of the three subjects, circles of slightly different sizes were used for each subject. However, relative displacements are not affected by this procedure. Since only the lower lip showed sizeable variations in all three subjects, parametric representations of the upper lip will not be considered.

Plots and transformations of the parametric data

Various kinds of plots and transformations have been attempted on the parametric data. Three that have proved useful will be discussed below.

1) Raw plots.-- These are simply plots of the raw measurements in scales proportional to the original scale. For these plots no speaker-dependent adjustments are made.

2) Centered plots.-- These are the same as the raw plots except that the units of each axis are measured as deviations from the means of the measures for each subject. One speaker-dependent measure per dimension is required.

3) Range-normalized plots.-- In these, each subject's score on a parameter is the position it occupies in the range of that parameter expressed as a percentage; i.e.,

$$N(x) = (x-min)/r$$

Where $N(x)$ is the range-normalized score, x is the raw parameter measurement for the subject in question on vowel [x]; min is his minimum measurement on that parameter; and r is his range for that parameter. The range normalization requires two pieces of speaker-dependent information per dimension.[15]

As indicated earlier, there seems to be little reason to object to raw parameter plots measured in the maxillary coordinate system to compare tongue positions because of the small variations in the size of the different subjects. Indeed, a centered plot of tongue circle positions proves to be little different from a raw plot. Different sizes of the mandible and lower lip make a centered plot desirable for cross-subject comparisons of lower lip position.

Generally more favorable results, in terms of the overall separation of phonetically distinct vowels, were found in the range-normalized plots of articulatory data for tongue position. However, there appears to be little justification for the use of such large amounts of speaker-dependent information given the overall similarity of the unnormalized acoustic data (see below).

Phonetic transcriptions

As noted above, there is a small amount of dialect variation among the three subjects. Some of the within-category variation in both the articulatory and acoustic events may be associated with these narrow phonetic differences. For this reason, an experienced phonetician, Dr. A. S. Abramson was asked to provide (moderately) narrow phonetic transcriptions of the experimental data. In order to facilitate the transcription process, a tape was dubbed from the tapes of the original experiment. In the dubbing process, the original tape was brought to the beginning of each experimental sentence three times, in succession, so that the dubbed tape contained three repetitions of each of the original sentences. The transcriber was given control of the tape recorder so that he could listen to any of the tokens as often as he wished.[16]

The transcriber was informed that the purpose of his transcriptions was to provide an independent check on possible dialect differences among the three subjects. His transcriptions for the three subjects are provided in Table 2.3.1.

TABLE 2.3.1

PHONETICIAN'S TRANSCRIPTIONS

Category	GNS	FSC	TMN
/bib/	i	$I^{\dot{}}$	$\overset{I}{\wedge}\overset{\ddot{}}{i}$
/bIb/	$I^{>}$	I	$I^{>\ddot{}}$
/beb/	$e^{\vee}i$	$e^{\dot{}}$	$e^{\dot{}}$
/bɛb/	$\varepsilon^{\wedge}$	$\varepsilon^{\wedge\ddot{}}$	$\varepsilon^{\wedge}$-I
/bæb/	$æ^{\wedge}$	$æ^{\wedge\ddot{}}$	æ
/bʌb/	$\Lambda^{<}$	$\Lambda^{<}$	ʌ
/bɑb/	$\alpha^{\ddot{}}$	$\alpha^{<}$	$a^{>}$
/bɔb/	$\mathfrak{o}^{\vee<}$	ɔ	$\mathfrak{o}^{\wedge\ddot{}}$
/bob/	$\Lambda^{>}\overset{u}{\wedge}$	$\Lambda^{>\vee}\overset{u}{\wedge}$	$o\overset{u}{\wedge}$
/bʊb/	U	$ɯ^{\vee}$	U
/bub/	u	$u^{<}$	$u^{<}$

Correspondence of traditional features to physical parameters

In the following sections, measures of tongue position will be compared to the traditional feature specifications of tongue height, tongue advancement and lip rounding. At the same time, correspondences of these features to the acoustic parameters F1 and F2 will be considered. The results generally corroborate the conclusion of Ladefoged et al. (1972) that acoustic parameters correspond better to traditional features than do articulatory measures. Furthermore, cross-speaker variation appears to be much less severe in the case of formant frequency measurements. In the articulatory parameter, large idiosyncratic differences are found within a single vowel category for different subjects.

Overall patterns of tongue positions and formant frequencies

Tongue positions.-- The pattern of tongue position as indicated in the combined raw tongue circle plot (Figure 2.3.5) seems to vary along a curve, from a position indicative of pharyngeal constrictions in /ɑ/ and /ɔ/ on roughly a 45 degree angle through /ʌ/, /o/ and /U/. It then continues nearly horizontally through /æ/, /ɛ/, /e/ and /I/. The vowels /i/ and /u/ appear to be off this main curve of variation. While there does appear to be a general pattern, there is considerable diversity in detail. The pattern displayed in the range-normalized tongue circle plot (Figure 2.3.6) is not substantially different from that described above.

The range-normalized average pellet plot (Figure 2.3.7) also shows a similar pattern. Perhaps the most noticeable difference is the relatively increased "backness" of /u/. This increased backness appears to be due to the greater bunching of /u/ relative to the other vowels.

Affinity structure anomalies in tongue position.-- The two range-normalized plots (Figures 2.3.6 and 2.3.7) display rough correspondence to some of the relationships posited by traditional analyses. There are, however, substantial departures in the details of the overall affinity structures of traditional analyses and any of these measures of tongue position. On the normalized tongue circle plot (Figure 2.3.6), for all three subjects, the vowel /æ/ is actually closer to /U/ than the latter is to /u/, its near neighbor in traditional analyses. This situation is only slightly improved on the normalized pellet plot (Figure 2.3.7). Traditional analyses usually describe tongue position of /u/ as "high back" and /U/ as a fronted and lowered "version" of /u/. The vowel /æ/ is generally given a diametrically opposed tongue-position specification, "low, front." Composite tracings of tongue contours for the vowels /u/, /U/ and /æ/ are provided in Figure 2.3.8 for each of the three subjects.

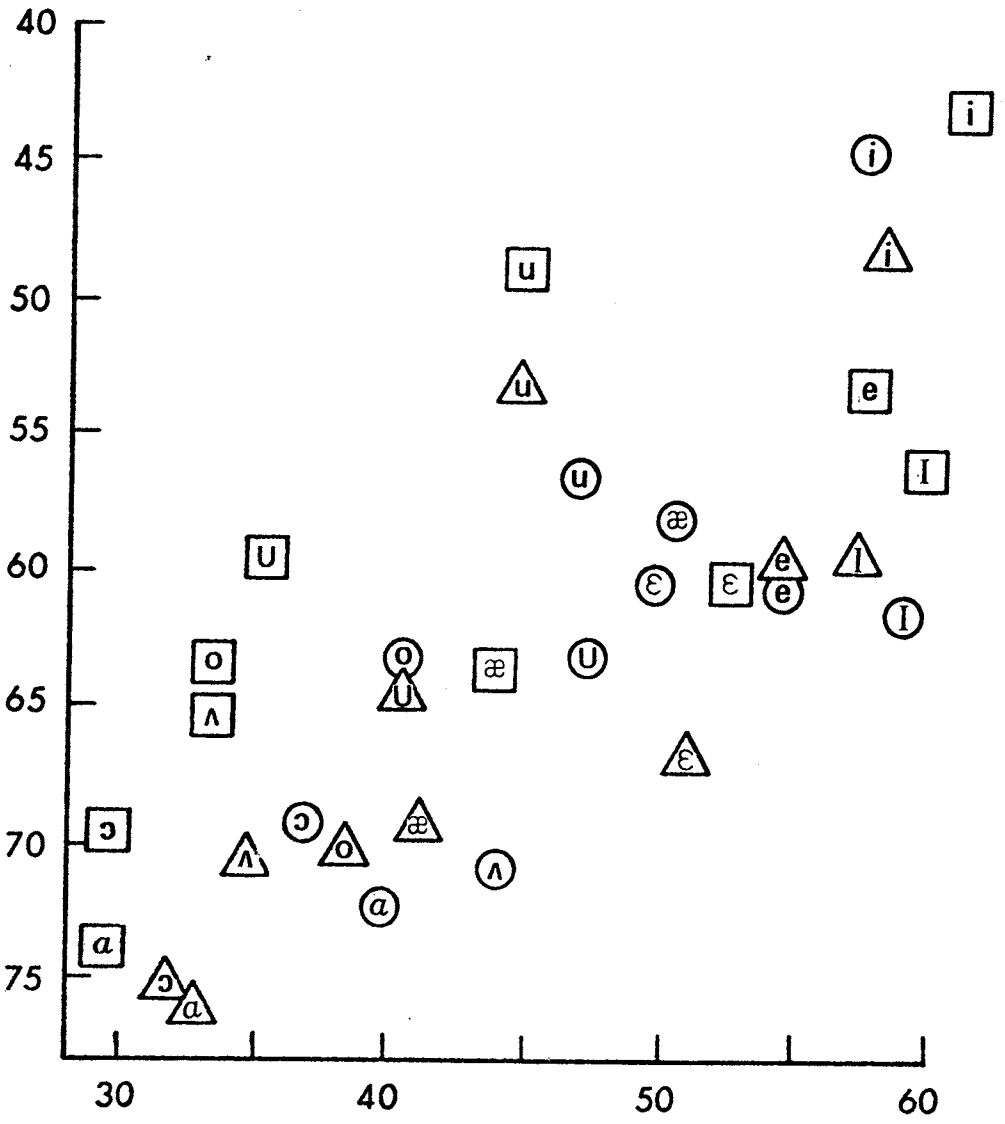


Figure 2.3.5. Raw tongue circle plot. Axes in millimeters from origin of maxillary system

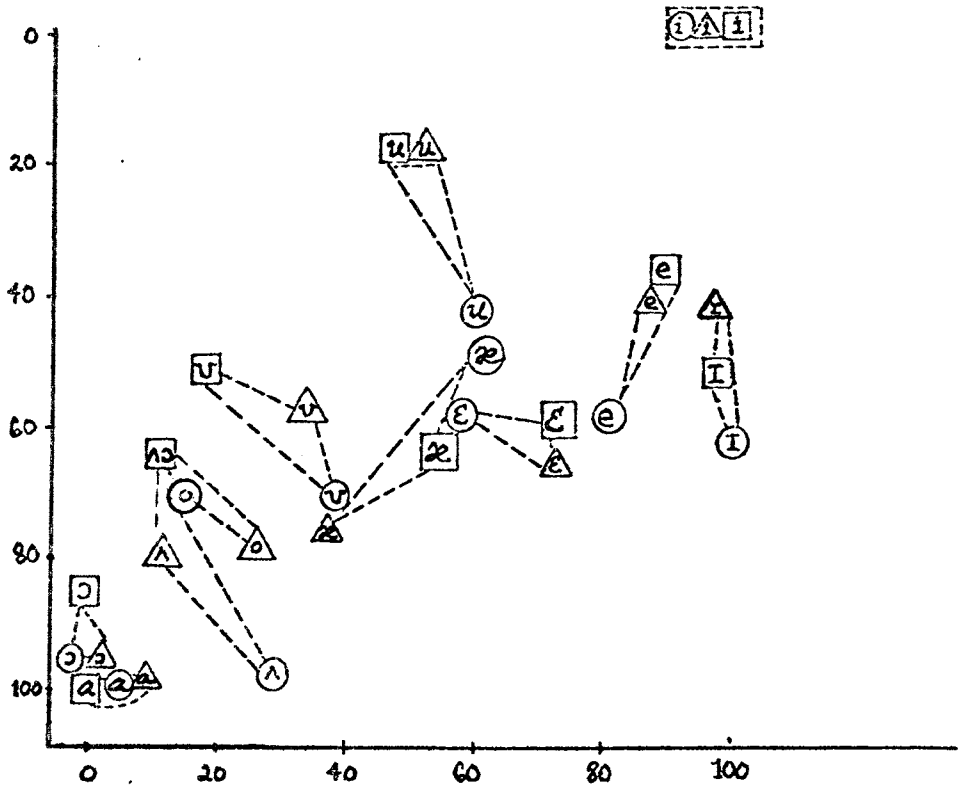


Figure 2.3.6. Range-normalized tongue circle.

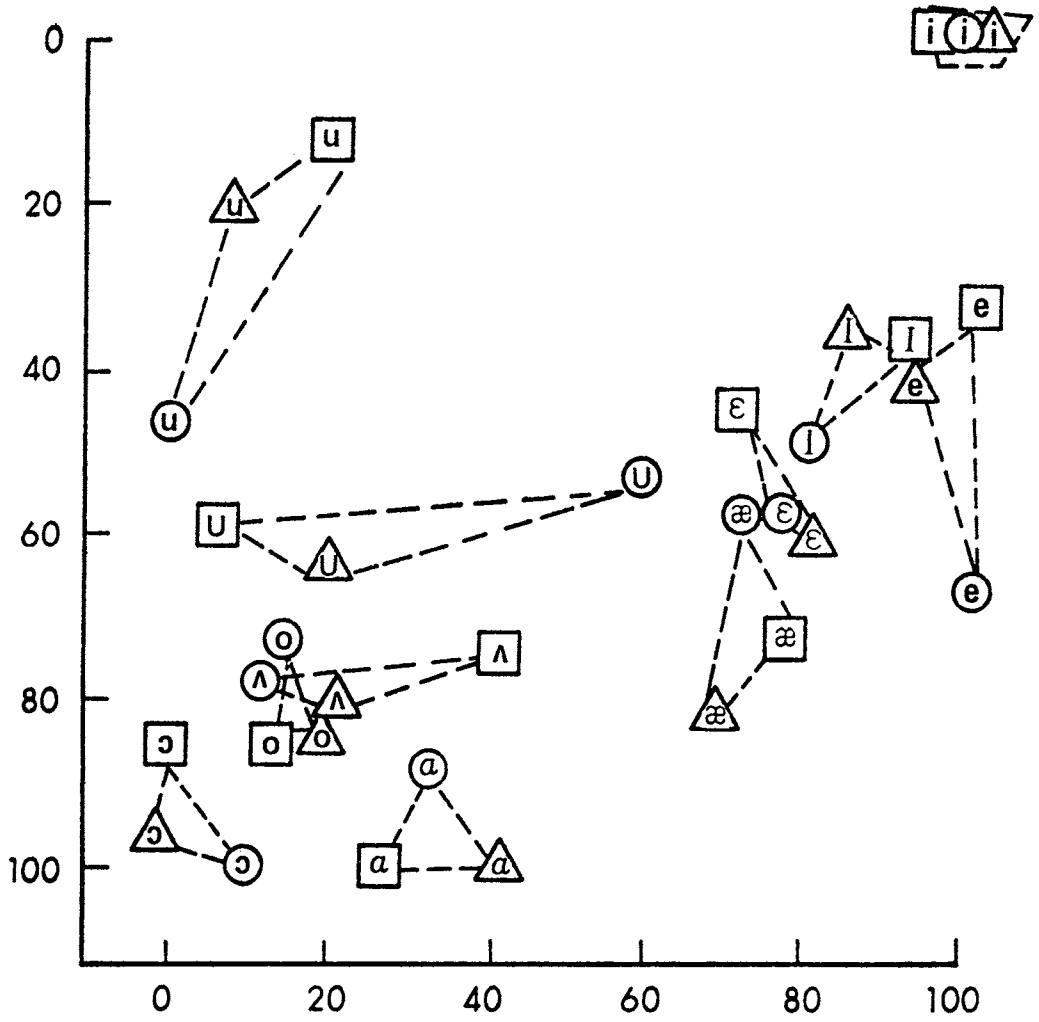


Figure 2.3.7. Range-normalized average pellet.

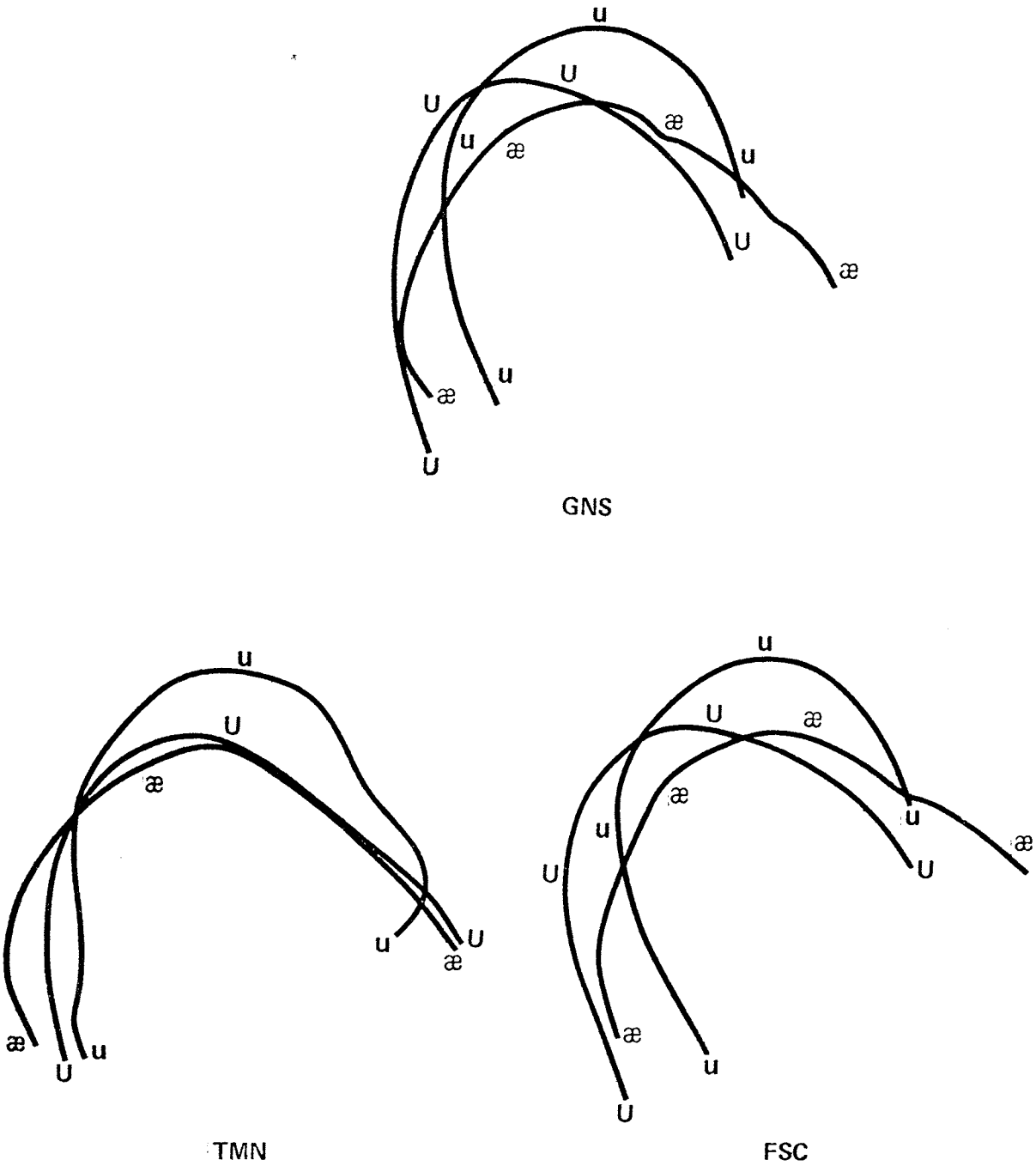


Figure 2.3.8. Tongue contours for [u], [U], [æ].

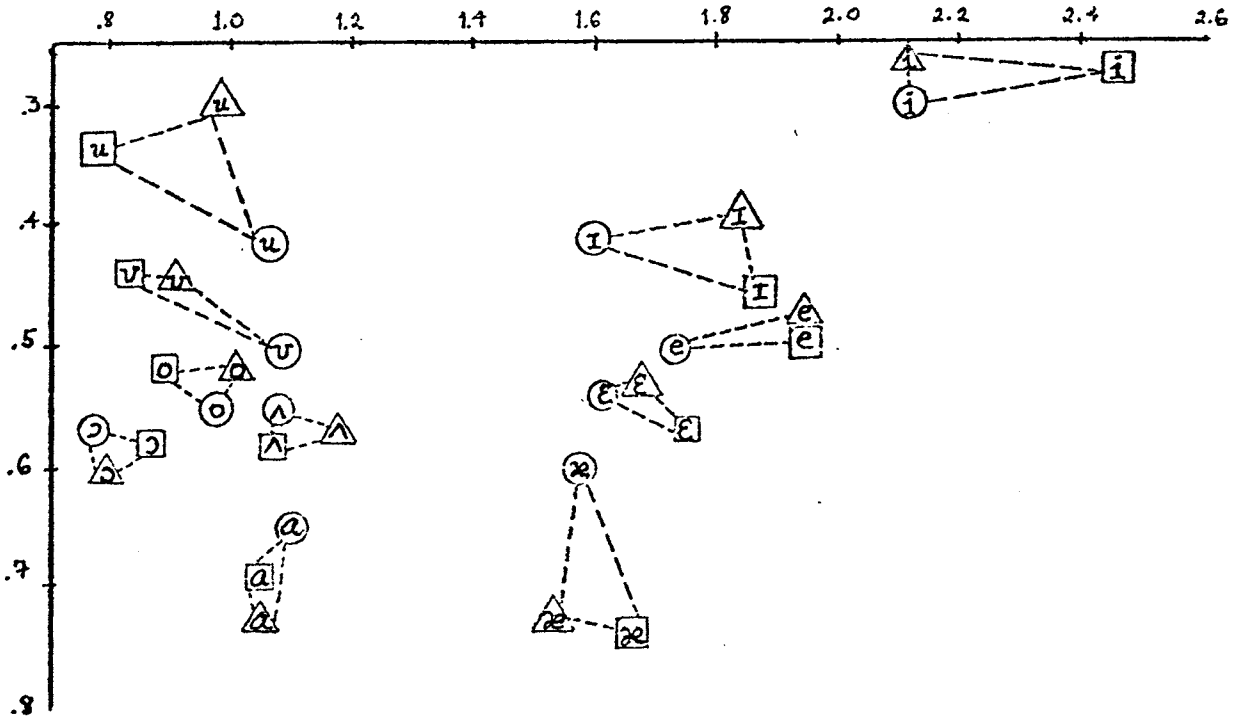


Figure 2.3.9. F2 by F1. Axes in kilohertz.

Acoustic pattern.-- A plot of the first and second formant frequencies of the three subjects (Figure 2.3.9) is in much better accord with the traditional vowel diagram (cf. Section 1.2 above).[17] In general, the vowels are well separated in the acoustic plot.

Though there is no overlap for vowels of traditionally different tongue positions in the range-normalized pellet plot (Figure 2.3.7), it should be remembered that the range-normalized articulatory plot requires two pieces of speaker-dependent information, while the acoustic plot is of raw formant measurements. From the graphs, it appears that there is substantially less speaker-dependent variation in the raw acoustic parameters than there is in the best plot of tongue positions found.[18]

As noted above, the tongue positions of /æ/ and /U/ are much more similar than traditional theory would have us believe. It is interesting to note that the vocal tract analog experiments of Stevens and House (1955) anticipate the possibility of such a result. It is apparent in their work (cf. their Figure 7, p. 41) that there exist possible tongue configurations such that formant frequencies near those of /æ/ can be changed to values near those for /U/ by a change in mouth opening (lip closure and protrusion) parameters only. From the composite tracings of these vowels (Figure 2.3.8) it would appear that for subject TMN this theoretical possibility is very nearly realized.

Further details of the articulatory and acoustic data will be considered in a feature-wise analysis of the correspondence to traditional theory.

Horizontal tongue position and the traditional advancement feature

If each subject is considered separately, it is possible, on any of the parametric representations of tongue position to separate the front vowels /i, I, e, ɛ, æ/ from the back vowels /u, U, o, ɔ, ɑ, ʌ/. However, when the three subjects are considered together, only in the normalized average pellet plot (Figure 2.3.7) can such a separation be made. Even in this plot, the vowel /U/ for FSC is closer to tokens of the front vowels /æ/ and /ɛ/ than it is to tokens of any back vowel, though it is recorded as a back vowel by the transcriber. (Table 2.3.1). The F1-F2 plot (Figure 2.3.9) shows a large separation (350 hertz minimum) in F2 between the front and back vowels. Thus though articulatory positions correspond to some degree to the traditional front-back distinction, the correspondence with F2 appears more clear-cut.

Vertical tongue position and the traditional height distinction

Height in front vowels.-- Figure 2.3.10 presents composite tracings for the front vowels of each subject. The rank order of height relationships appears to be in accord with traditional statements

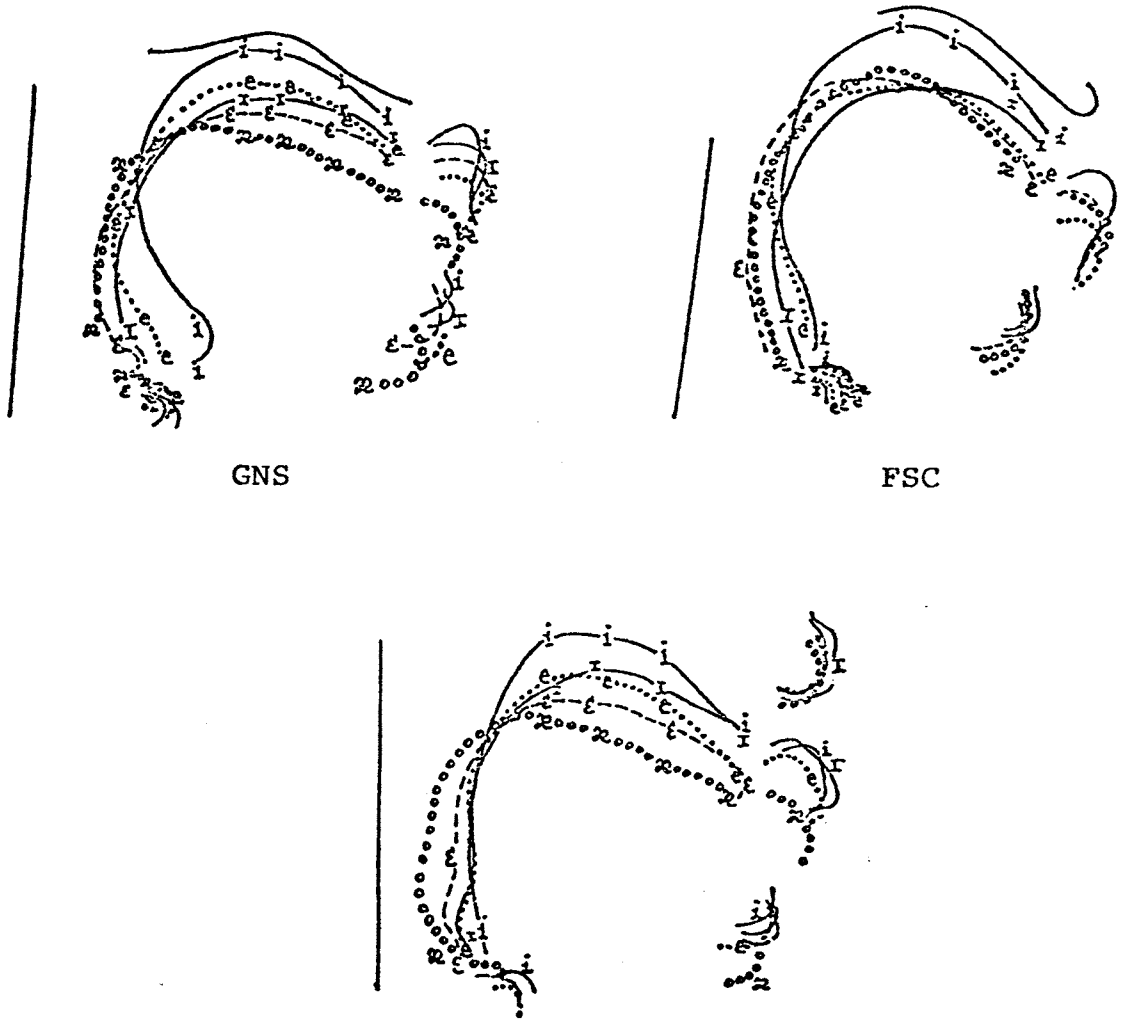


Figure 2.3.10. Composite tracings for front vowels.

[i] ————————
[I] ————————
[e] . . . . . .
[ɛ] — — — —
[æ] o o o o o o

only for subject TMN. Even in this case, the height relationship between /e/ and /I/ is somewhat ambiguous in that the height at the highest point of the tongue is the same for these two vowels. Subject GNS shows a rather clear "inversion" of the traditional order for this pair.

There are some objections that may be raised here because of the diphthongal character of /e/. Though considerable care was taken for this vowel to ensure that both the acoustic and articulatory records were at points of minimal change, some doubt must remain as to the success of this procedure in these cases of especially critical timing. Notice, however, that the acoustic data (Figure 2.3.9) for all three subjects is in accord with the usual traditional height assignments. These articulatory observations are consistent with findings noted earlier by Ladefoged et al. (1972), Parmenter and Treviño (1932) and Russell (1928).

A more serious problem of correspondence of tongue position to traditional height specifications is evident in Figure 2.3.10 for the front vowels of FSC. The highest point of FSC's /æ/ is actually higher than that of his /I/.[19] Interestingly, this appears to corroborate a claim of Russell (1928) quoted in Section 2.2 above. From the composite plot (Figure 2.3.10) it appears that the articulatory distinctions are made almost exclusively outside the oral cavity: in the pharynx for /e/ versus /ɛ, æ/ and at the lips (and perhaps marginally by different larynx heights as indicated indirectly by hyoid height) for /ɛ/ versus /æ/.

The problems described above are further confounded when parametric representations including data from the three subjects together are considered (Figures 2.3.5, 2.3.6 and 2.3.7). There is overlap among vowels of different classes along the vertical tongue position coordinate even in the range-normalized average pellet plot (Figure 2.3.7).

In the acoustic plot (Figure 2.3.9), FSC shows a somewhat lower F1 for /æ/. However, the traditional feature relationships appear to be reasonably well represented in the F1-F2 space. It is possible to distinguish the front vowels in this plot by drawing boundary lines perpendicular to the F1 axis. That is, there is no F1-overlap among the front vowels of different categories.

Height in back vowels.--

Figure 2.3.11 presents composite tracings for the back vowels of each subject. Ordinal height relationships are generally in better agreement with traditional specifications than for the front vowels.[20]

There are, however, problems for height correspondence here as well. Perhaps the most striking facts to be noted in the

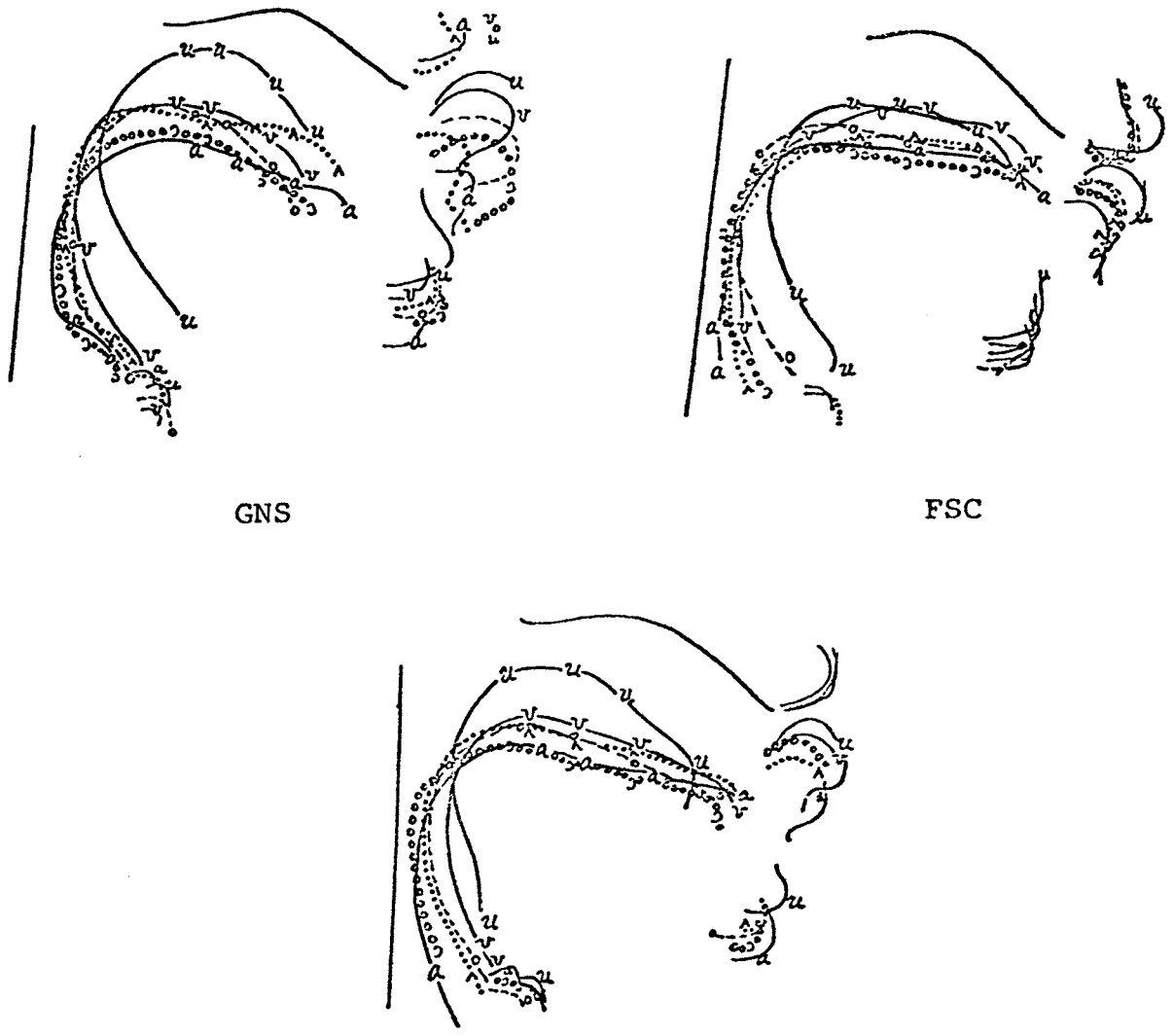


Figure 2.3.11. Composite tracings for back vowels.

[u] ————————
[U] ————————
[o] — — — — —
[ɔ] ° ° ° ° °
[ʌ] · · · · ·
[ɑ] ————————

composite tracings (Figure 2.3.10) are 1) the large contrast in the tongue contours in the mouth for /u/ versus the other back vowels in subjects GNS and TMN (a similar remark applies to FSC's tongue contours in the pharynx); and 2) a general lack of contrast in tongue contours in the mouth region among vowels of traditionally different heights. For all three subjects, though the main variations in vocal tract configurations distinguishing the vowels are distributed in different parts of the vocal tract, the articulatory differences separating /U/ from its near phonetic neighbor /u/ are more striking than the differences separating any of the rest of the vowels from each other, including /U/ and /ɑ/. The acoustic record, on the other hand, shows differences which are more in accord with the implied distance relationships of traditional phonetics.

Parametric plots of articulatory data do not appear to provide substantially better correspondences in tongue height to the relative height categorizations of traditional phonetics. However, the normalized pellet plot (Figure 2.3.7) does provide reasonably good separation of /u/ versus /U/ versus /o,ʌ/ versus /ɔ,ɑ/. In the acoustic plot (Figure 2.3.9) it is possible to separate all of the vowels along the back line of the vowel triangle (i.e. /u, U, o, ɔ, ɑ/) on the basis of F1 alone. The vowel /ʌ/ overlaps with both /o/ and /ɔ/ in F1.

Correspondence of "height" parameters in front and back vowels.--The problem of lack of parallelism of the articulatory correlates of the traditional height feature in front and back vowels has been discussed briefly by Delattre (1951) and by Ladefoged (1967, 1974). The lack of such parallelism appears to be a prime factor in Ladefoged's rejection of "tongue height" as the defining parameter of the feature "vowel height" and for his adoption of an F1 related definition in its place (Ladefoged 1971a). Delattre appears to suggest that a rotation of the articulatory figure might provide a better correspondence of height parameters for front and back vowels. No rotations would appear to do much good in any of the articulatory plots of data presented here. Any rotation of axis to attempt, for example, to equalize values on the vertical coordinate for /i/ and /u/ for any subject would have detrimental effects elsewhere for height correspondence.

Acoustic plots and perceptual experiments, as Ladefoged and Delattre have pointed out, generally show that vowels of the same traditional "height" classifications have nearly the same F1 for both back and front vowels. Our present acoustic plot (Figure 2.3.9) corroborates this finding for these subjects. There is some evidence of a slight rotation of the correspondence of F1 relationships to traditional height classes in that the average values of the F1's of back vowels are consistently slightly

higher than those of corresponding front vowels of the same traditional height classification.

Lip rounding

Lower lip position as measured in the maxillary coordinate system (by fitted "lip circle") is the result of both jaw position and displacement of the lips relative to the jaw. The horizontal coordinate of lip position will be referred to as *extension* and the vertical coordinate, *elevation*. Extension is to be distinguished from *protrusion* in that the latter will be defined as the component of extension which is due to (presumably active) lip displacement beyond the mandible.[21] The action of closing the jaw results in lip elevation and lip extension, but not in lip protrusion. Lateral spreading is another lip gesture of possible interest, but since it is generally not available in mid-sagittal X-ray views it cannot be considered here. The term lip position as used here refers to the extension and elevation of the lips only.

General correspondence to traditional specifications.-- In the combined three subject centered plot of lip position (Figure 2.3.12), it is possible to draw a boundary between the generally rounded vowels /u, U, o, ɔ/ and the rest. However, the vowel /U/ of subject FSC is transcribed as an unrounded vowel in the [ɯ]-[ɣ] area by the trained listener. The centered lip position plot for FSC alone (Figure 2.3.14) indicates that his lip position for /U/ patterns with the vowels /ɔ, o, u/, all transcribed as rounded vowels. However, the acoustic parameters provide evidence for an auditory basis for the difference in rounding transcription for this vowel. From Figure 2.3.9, it appears that FSC's /U/ has a relatively higher F2 than any of the vowels that are transcribed as back rounded. Furthermore, the relative position of this vowel in the F2 space is roughly analogous to one intermediate between that of [u] and [ɣ] in the 1910 IPA vowel diagram. (See Chapter I, Figure 1.2.3.)

The discrepancy between lip position and the phonetician's transcription is limited to this one token in the present data. However, the fact that the discrepancy appears to have acoustic motivation, taken together with the evidence for the relative variability of phonetician's transcription of rounding (Ladefoged 1960, Laver 1965) casts further doubt on rounding as an independent, acoustically recoverable feature. The possibility remains that a single F2-based phonetic feature, "clarity", might underlie the transcription practice of phoneticians. The fact that synthetic two-formant continua produce the impression of both "rounded" and "unrounded" vowels is consistent with this notion. (Cf. Delattre, Liberman and Cooper 1951) Under this interpretation, rounded

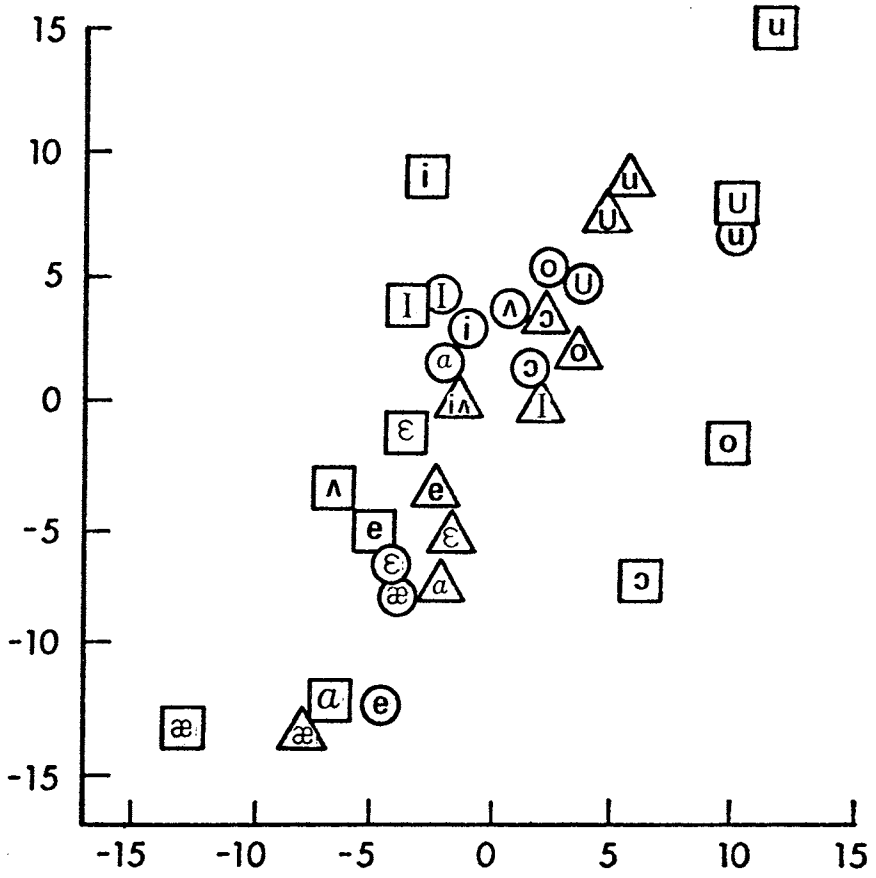


Figure 2.3.12. Centered lip-circle plot for three subjects.

vowels are simply "darker" (having lower F2 values) versions of the unrounded vowels with the same traditional advancement and height specifications.

Idiosyncratic strategies.-- If, on the other hand, rounding is a feature, it is difficult to see how it is to be specified in articulatory terms. The individual lip position plots (Figures 2.3.13-15) make it clear that there are considerable differences among subjects in their general strategies of lip use. Subject GNS (Figure 2.3.14) shows a relatively large difference in the extension of the lips for rounded and unrounded vowels at roughly the same elevations. It is evident from the composite tracings of the back vowels (Figure 2.3.12) for this subject that he makes extensive use of lip protrusion.[22] GNS's vowels appear to be relatively evenly distributed along the elevation dimension.

FSC's pattern is quite different (Figure 2.3.14). While extension of the rounded series is greater than that of the unrounded, the differences are more gradual. On the elevation dimension, FSC shows two clusters with relatively less elevation for the vowels /ɛ, æ, ɐ/ and relatively more for the others.

The pattern of lip positions for subject TMN is different from either of the others, the strong correlation evident in Figure 2.3.15 between extension and elevation suggests that he makes very little use of protrusion but relies on jaw opening to produce different mouth openings. Lack of protrusion is generally confirmed in the composite tracings of Figure 2.3.11 for this subject.

Acoustic considerations.-- Thus while there is certainly a tendency for more extended and elevated lip positions for vowels traditionally described as rounded, the individual variation is quite striking. The complex patterns of lip position tend to suggest that it is not a physical manifestation of an independent phonetic feature. However, acoustically, lip positioning may be viewed as the control of a variable impedence at the end of the vocal tract (Stevens and House 1961) that interacts with other aspects of the vocal tract shape to determine the transfer function. If articulatory activity in vowels is viewed as the tuning of a complex acoustic network for the production of acoustically definable contrasts, the variability in strategies of lip use by different subjects is not surprising. For if the contrasts are so defined, how a subject manages to produce them may be of no importance to anyone except the subject himself who has virtually unlimited time to develop and adopt strategies to control the unique speech synthesizer that is his own vocal tract.

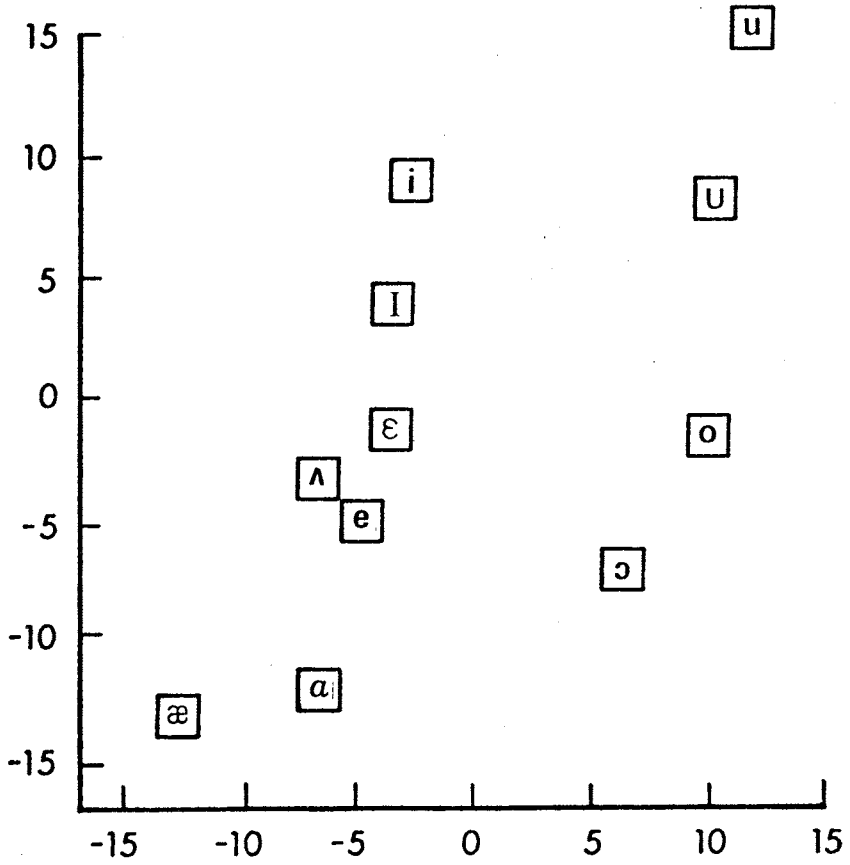


Figure 2.3.13. Centered lip circle plot for subject GNS. Axes in millimeters from subject's average.

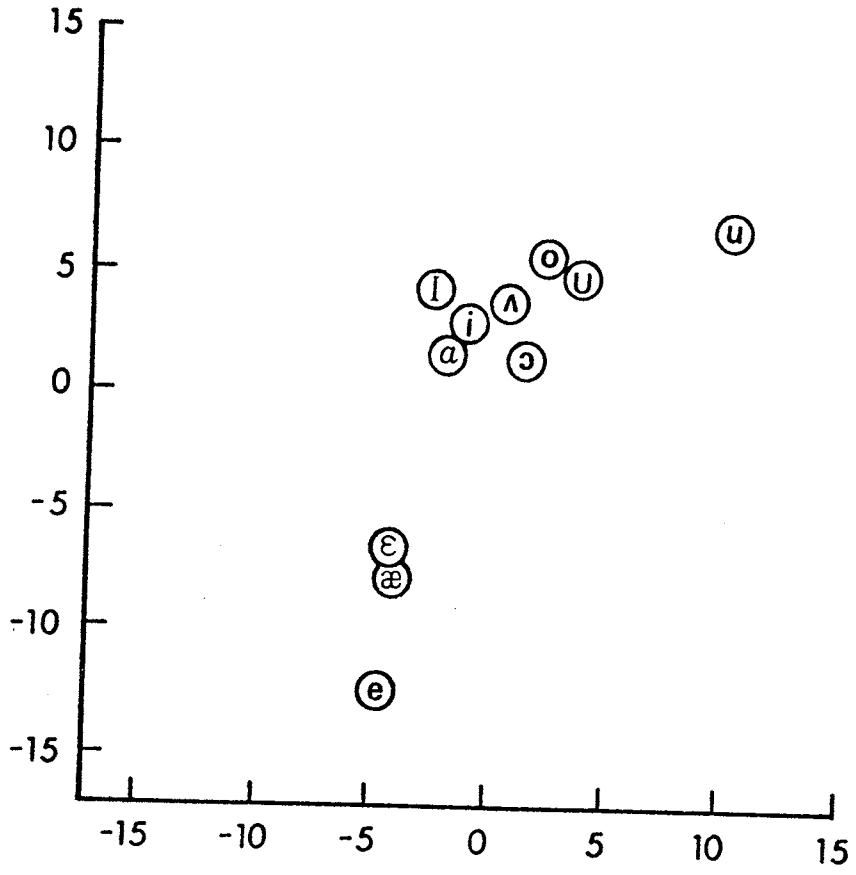


Figure 2.3.14. Centered lip-circle plot for subject FSC. Axes in millimeters from subject's average.

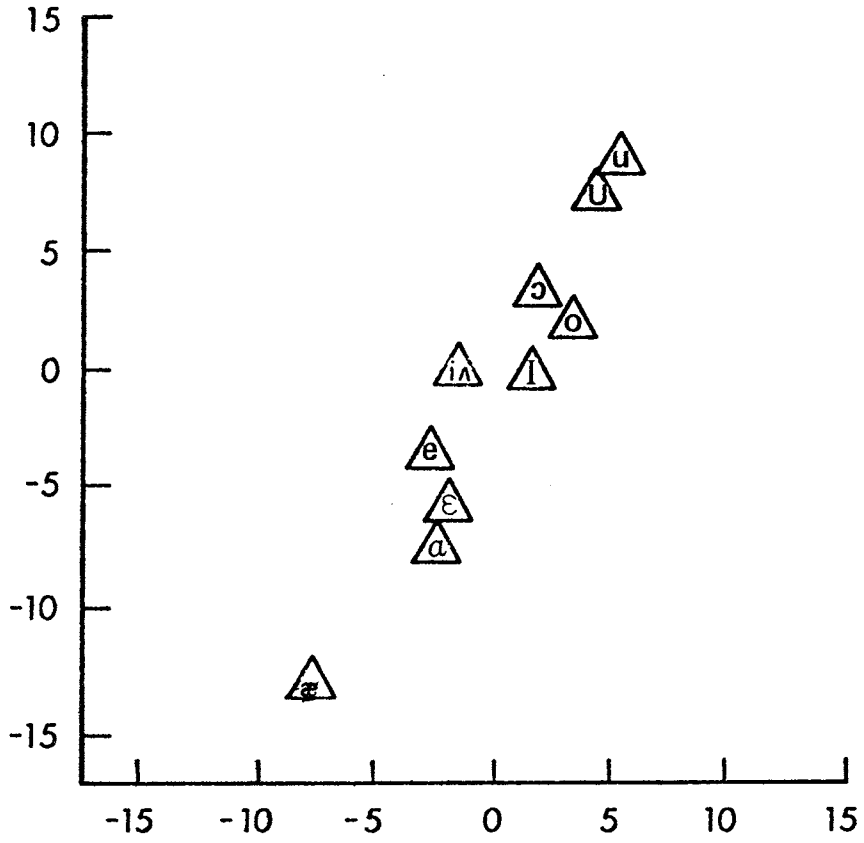


Figure 2.3.15. Centered lip-circle plot for subject TMN. Axes in millimeters from subject's average.

Tense-lax and tongue shape

As mentioned in section 2.2, the features "advanced tongue root" (ATR) and "constricted pharynx" (CPH) have recently been suggested as a feature complex to replace the (somewhat controversial) distinction between tense and lax vowels. Ladefoged et al. (1972) have suggested that some general "tongue bunching" parameter might better correspond to the distinction in question, though their investigation indicated that tongue shape was not consistently associated with tense-lax distinctions.

In the present study, no explicit measure of tongue root position was attempted since the location of this structure cannot generally be determined with any precision in the single frames. However, an examination of composite tracings tends to corroborate recent criticism of tongue shape features. The front and back series of vowels will be discussed separately.

Shape in front vowels.-- According to Perkell (1971), the vowels /i, e/ are "tense" because they possess the feature +ATR. The vowel /æ/ is "tense" because it is +CPH. The vowels /I/ and /ɛ/ are "lax" because they are specified -ATR AND -CPH. All three subjects in the present study do show a more advanced tongue position at the part of the tongue near the hyoid for /i/ vs /I/ and for /e/ versus /ɛ/. But TMN shows nearly identical pharyngeal configurations for "tense" /e/ and "lax" /I/. (See Figure 2.3.10.) On the other hand, ONLY subject TMN shows a more constricted pharynx for the vowel /æ/ than for the other front vowels. Both FSC and GNS show nearly identical tongue contours in the pharynx for /æ/ and /ɛ/.

Shape in back vowels.-- The situation for the back vowels with respect to the new feature is hardly better than for the front. The "tense" vowels /u/ and /o/ are analysed by Perkell as +ATR and the tense vowels /ɔ/ and /ɑ/, as +CPH. The lax vowels /U/ and /ʌ/ are analysed by Perkell as -ATR, -CPH. Though /u/ has a consistently more expanded pharynx than /U/ (and in fact more than any other back vowel), this is not the case for /o/ versus /ʌ/. See Figure 2.3.11. Only subject FSC shows any sizeable differences of the tongue contours in the predicted direction for the pair /o/-/ʌ/; both subjects TMN and GNS show the lax vowel /U/ to have more ATR than /o/. The feature "constricted pharynx" might be maintained for all three subjects for the vowels /ɑ/ and /ɔ/, though these vowels appear to be more constricted to different degrees in different places for the different subjects.

Optimal circular bands and tongue shape.-- Rough indication of the relative bunching of the tongue can be had in a plot of the

optimal circular band data shown in Figure 2.3.16. The X-axis represents the measure in units of five degrees of arc of tongue contour contained in the optimal circular band. A higher score on the X-axis will be called a "more circular" tongue shape. The Y-axis represents the diameter of the circular band. A tongue shape fitted to a larger circular band will be called "broader" and a smaller one "tighter-arched." In general, more circular and tighter-arched shapes correspond to the intuitive notion of "more bunched" articulations, but there is no simple interpretation when two indices point in opposite directions. For all three subjects, within subjects, the vowels /i/ and /e/ have either "more circular" or "tighter-arched" shapes than the corresponding /I/ and /ɛ/. For all subjects /u/ shows a more circular shape than /U/, though /U/ shows a tighter arch for TMN and GNS. No other generalizations related to any traditional features or modern modifications thereof appear in this parametric representation of tongue shape.

While the vowels /i/, /ɑ/, /ɔ/ and /u/ tend to show more extreme shapes, there is generally very poor correspondence between tongue features and the purported feature "tense-lax". While no articulatory features appear to be represented as clearly in articulatory data with the consistency and regularity that would seem to be demanded if the feature has its basis in the articulatory property in question, "tense-lax" distinctions seem to be particularly poorly related to any apparent articulatory parameter. Perhaps this accounts for its marginal status in traditional feature theory.

Conclusions

While there appears to be some general relationship between the traditional features advancement, height and to a lesser degree rounding, the relationship between the affinity structure of traditional phonetic descriptions and F1-F2 plots appears more satisfactory. Furthermore, articulatory configurations appear to contain large speaker-dependent components which are not manifested in either the acoustic output or in the phonetic transcription. As a theory of the relationships between physical parameters and presumed phonetic structures, traditional statements about articulatory configurations appear to be inadequate. However, even the rather imprecise correspondence between traditional descriptions and general patterns in articulation may render them an invaluable aid in the teaching of practical phonetics. On the other hand, at least in the special case of subjects with roughly the same vocal tract size, the most salient aspects of the traditional affinity structure for vowels are more directly relatable to information in the first and second formant frequencies.

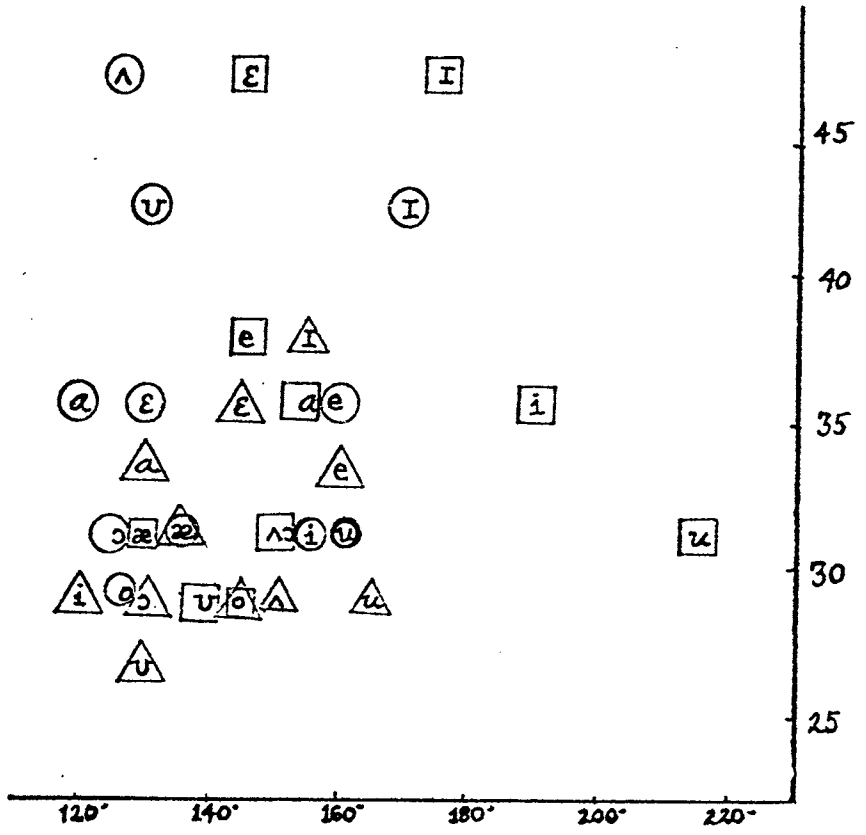


Figure 2.3.16. Optimal circular band plot. Vertical axis is radius in millimeters of optimal circular band. Horizontal axis is degrees of arc of tongue contour contained within optimal circular band.

More complicated cases of acoustic to phonetic correspondence, where the absolute frequency values of different speakers' vowel formants vary considerably, are considered in the next two chapters.

NOTES TO CHAPTER II

[1]The term "phoneme" as it is used by authors cited in this section does not correspond to any of the phonological units bearing the same name posited by various linguistic schools of this century. Rather, the term appears to correspond more closely to the traditional notion of phone or speech sound, or, perhaps, to the recently discussed concept "extrinsic allophone" (cf. Tatham 1971 and references listed there). MacNeilage notes that the phoneme was chosen as the basic unit of description only as "... a descriptive convenience and was not meant to imply that there was only one target per phoneme"(1970:192).

[2]All of the passages from Russell (1928) quoted below have been cited in footnotes by Parmenter and Treviño (1932).

[3]The Carmody study also carefully controlled head positions.

[4]The defense of traditional theory appears to have been motivated primarily by the fact that there was nothing available to supplant it. Carmody remarks:

> ... for the language teacher, our films come to the aid of the much slandered vowel triangle. Molded into a quadrilateral (with one problematic angle) and modified by simultaneous changes of the lips and pharynx, the principle of the triangle is far sounder than are the spectrum, rectangular, dull and bright and other similarly subjective classifications (1937:vi).

It is doubtful that the term "spectrum" in the above passage refers to actual spectral analyses of speech, since only very few such analyses were available at the time. Russell himself, in rejecting traditional theory, provided no coherent alternative. Rather, he seemed to advocate almost any substitute for the articulatory vowel triangle. His flirtation with several off-beat theories of speech production (together with his flamboyant and frequently less than lucid style) may have seriously weakened the impact of his substantial empirical contribution.

[5]Carmody provides only a brief comment on Holbrook's phonetic notation. From this account, it appears that the grave (ˋ) is meant to indicate generally less peripheral (more centralized) vowel qualities; and the acute (ˊ), more peripheral. The other diacritics appear to be used in a similar fashion to those of the IPA.

Notes

[6]Perkell (1965:39) suggests a "synergy" between lip and laryngeal mechanisms as part of a vocal tract length changing system. Were the tradeoff solely between the lip protrusion and larynx height, an articulatory feature of vocal tract length might be proposed. However, the concomitant variation of lip CLOSURE would seem to point strongly to a purely acoustic motivation to the various gestures.

[7]The repetition stability of articulatory configurations would appear to detract from an extreme version of this view. Since naive subjects spontaneously assumed nearly identical articulatory postures for the same vowel, even after periods of weeks or months between repetitions, it seems likely that sustained pronunciations are relatively natural extensions of ordinary speech rather than ad hoc gestures.

[8]Remarks about the failure of traditional theory to account for large variations in the pharyngeal region had been noted already by Russell (1928). Fant (1960) also stresses this oversight. With the exception of experimentalists, until recently phoneticians have generally considered the changes in pharyngeal region as "secondary" articulations, analogous to nasality, particularly in the description of "pharyngealized" consonants in languages such as Arabic.

[9]Lieberman (1976) cites this evidence as well as earlier observations of Perkell (1965) as reason for rejecting features such as ATR which imply consistent local deformations of the tongue.

[10]The broad transcription in this case corresponds roughly to the level of "taxonomic phonemes" or to the contrasting elements of the "surface phonology" of English. Though there are substantial difficulties in the formal definition and phonological justification of such a level (Chomsky 1964), more abstract (morphophonemic) representations are of little use in experimental phonetics.

[11]Most of the soft palate thus follows roughly the X-axis of the maxillary coordinate system described below.

[12]Thisvalue is much smaller than differences in head position found by Parmenter and Treviño (1932) to cause large changes in the vowel articulations of a single subject.

[13]Thus for subject FSC, the total range of the horizontal motion of the lower lip pellet is only about 5/8 of the range of the "lip circle" measurements described below.

[14]No direct measure of tongue pass has been attempted, since it appears that there is no satisfactory method of determining if that is not extremely sensitive to small local variations.

Furthermore, determination of a region of maximum constriction is essentially impossible when it occurs near the soft palate, since this structure could not be precisely determined in the single frames.

[15]The range normalization procedure used here is based on the Gerstman normalization procedure discussed in the next chapter. The centered plots are parallel to additive point normalization procedures also discussed there.

[16]One of the sentences of subject FSC was inadvertently omitted in the dubbing process. It was transcribed by Dr. Abramson at the end of the session from the original experimental tape.

[17]It is also in better agreement with Singh and Woods' (1971) plot of the first two factors in their multidimensional scaling analyses of similarity judgments by naive English listeners.

[18]Indeed, the range-normalized average pellet plot has only been included because it provides the best separation of vowels of any method attempted. It is difficult to imagine what articulatorily invariant properties it could represent.

[19]Though FSC's /æ/ is transcribed with a centering offglide (see Table 2.3.1), it occurs later in the syllable and does not affect the measurement. The formants in the region of measurement show essentially no change and there is a "held" position of several frames in the articulatory data.

[20]This is somewhat surprising in light of statements of Russell (1928) and the comments on the French back vowels by Carmody (1937) quoted earlier. Ladefoged (1967, 1974) has also remarked on particular difficulties with height correspondence in back vowels to traditional statements in connection with published radiographs of a full set of cardinal vowels published by S. Jones.

[21]Similarly, lip closure could be defined as the component of elevation due to active vertical displacement of the lip with respect to the mandible. However, the distinction is not required in the ensuing discussion.

[22]It should be noted that the degree of lip protrusion is probably somewhat exagerated by increased radial distortion introduced by the cineradiographic process. While no measurement of this distortion was made in the current experiment, measurements provided by Kuehn (1973:35) using the same

radiographic equipment (at the Eastman Dental Center) indicate that it is probably negligible for all measures discussed here except in the area of the lips. However, relative positions of the lips shold not be seriously affected, since local comparisons involve roughly the same amount of radial distortion.

CHAPTER III

RELATIVE FORMANT NORMALIZATION AND THE PERCEPTION OF VOWEL QUALITY

3.1 The Problem of Cross-speaker Within-phone Variation

Introduction

In the last chapter, it was indicated that the acoustic parameters F1 and F2 are better related to phonetic properties of vowels than are the traditional articulatory parameters. However, even in the case of stressed vowels, there are serious problems of cross-speaker within-phone variation that must be faced by any explicit theory of vowel quality features.

Instantaneous F1-F2 measurements.-- Thusfar, nothing in the literature seems to contradict Joos' tentative conclusion that the frequencies of the first two formants "... may justifiably be called the principal determinants of vowel color!"(1948:65) In the rest of this work we will be limiting discussion of acoustic parameters almost exclusively to the frequencies of the first two formants. Other factors such as fundamental frequency and frequency of the third formant have been shown (under certain circumstances) to have some effect on the phonetic quality of vowels (Potter and Steinberg 1950, Miller 1953, Fant 1959, Carlson et al. 1970). However, the role attributed to these other factors has generally been that of producing a secondary modification of a basic F1-F2 pattern.[1] As in the previous chapter, we will ignore the segmentation problem and assume (at least in the case of stressed vowels) that a quasi-instantaneous measurement of formant frequencies contains all the relevant information in a syllable necessary to the specification of phonetic quality.

Cross-speaker variation in formant measurements.-- Joos was among the first to point out what remains one of the most difficult problems for the mapping of the values of F1 and F2 to phonetic features for vowels. While raw acoustic measurements appear to be reasonably invariant properties of vowels among speakers of roughly the same "head size", differences between the formant frequencies for the same vowel produced by men and children, according to Joos, "... are nothing short of enormous -- they are commonly as much as seven semitones, or a frequency ratio of three to two, about the distance from /ɛ/ to /U/ ..." (1948:65)

Joos makes three important points in discussing this problem. 1) The size of the variation between subjects is considerably larger than variations that normally occur within subjects on repetition. While he offers little more than anecdotal evidence in support of this claim, it is generally corroborated by Pottern and Steinberg (1950). 2) If vowels with formant frequency differences of the general magnitude and direction as those observed between the speech of men and children occurred within the speech of a single speaker, they would be phonetically distinct. That is, within-category differences between speakers produce overlaps in F1-F2 space. While Joos provides limited evidence supporting this claim, more striking examples of such overlap have been provided by Potter and Steinberg (1950) and Peterson and Barney (1952). 3) The same differences in formant frequencies regularly occur among speakers of the same dialect in phonetically equivalent items, "... without any one of them noticing anything peculiar about another's choice of vowel color."(Joos 1948:60)

To resolve these difficulties, Joos suggests that it is the RELATIONSHIPS among the formants of the vowels of a given speaker that remain invariant in spite of the variations in the absolute frequencies that occur across speakers. We will consider several attempts at the characterization of such invariant relationships.

Normalization procedures as feature extractors

It was argued in Chapter I that a primary goal of experimental phonetics is to formulate testable hypotheses of the mapping that exists between physical parameters and phonetic features. A function which attempts to make explicit a part of such a mapping will be referred to as a *feature extractor*.

The main task of such a feature extractor is to separate variation in physical parameters into two components: that which is associated with phonetic variation and that which is not. A class of algorithms known as (speaker-)normalization procedures can be viewed as putative feature extractors for vowel data. These attempt to eliminate the effects of speaker differences in phonetically equivalent events by "modifying" formant measurements on the basis of other information extracted from the signal.

It is assumed that phonetic feature specifications can be "read off" the normalized formant values. Such a procedure may be viewed as a system that makes explicit certain invariant relationships inherent in the signal.

There are two questions we will ask of a normalizaiton system. 1) To what extent do the relationships implied by the

normalization system actually occur in natural data; that is, to what extent can actual formant data be successfully normalized by the procedure in question? 2) Is human vowel perception sensitive to the relationships implied by the normalization system? Question 1 will be explored briefly in this chapter and in more detail in Chapter IV. Question 2 will be examined below.

Relative formant normalization.-- While there have been a variety of hypotheses concerning what additional information should be used to "modify" or "interpret" the F1 and F2 measurements, consideration will here be limited to RELATIVE FORMANT procedures. In normalizing the formant values of a particular vowel, these procedures use only information from the F1-F2 measurements of the vowel SYSTEM produced by the same speaker.

Range and point normalization.-- In the discussion of experimental indications of the relevance of relative formant information to perceptual evidence, we will distinguish between two types of relative formant normalization that have been proposed (at least implicitly) in the literature. These differ primarily in the amount of information about a speaker's vowel system that is necessary for the operation of the algorithm. In RANGE NORMALIZATION, the algorithm requires the formant frequencies of at least two vowels of known phonetic value. In POINT NORMALIZATION, only one known point in a speaker's system is required.

In order to provide a basis for the explicit comparison of different normalization procedures, a notational framework must be provided.

Notational conventions

Formants, labelled *F*, will be provided with up to three subscripts: for number (first or second formant), vowel and subject. The formant subscript will appear immediately after the symbol *F*, and a capital *N* or *M* will occasionally be used to substitute for constant formant numbers. Vowel subscripts will be placed in square brackets after the formant subscript and subject subscripts will follow these. Thus, $F_{N[V]s}$ is the *N*-th formant of vowel [v] for subject *s*.

Another convention that will prove useful from time to time is the dot convention used to indicate averaging over a particular subscript. Thus $F_{1[V]\cdot}$ would indicate the value of the first formant for the vowel [v] averaged over all subjects in question. $F_{M[\cdot]s}$ would indicate the value of the *M*-th formant for subject *s* averaged over all his vowels. We can also indicate averaging over two subscripts;

thus $F_{\cdot[\cdot]s}$ is the average formant value of both F1 and F2 for subject *s*. It is equal to $(F_{1[\cdot]\cdot} + F_{2[\cdot]\cdot}) / 2$. The overall mean for all subjects for both formants of all vowels is indicated by $F_{\cdot[\cdot]\cdot}$.

A normalized formant value will be indicated by an asterisk following the symbol F: F*; thus $F^*_{N[V]s}$ is a normalized value for the *N*-th formant for vowel *v* of subject *s*. In a perfect normalization procedure, for phonetically equivalent vowels spoken by different subjects, normalized formant values for corresponding formants of corresponding vowels should be equal. That is, $F^*_{N[V]s} = F^*_{N[V]t}$, for all *s* and *t*.

Linear rescaling and range normalization

Range normalization procedures imply that all the important information about relational invariances in a speaker's vowel system is determined by the relative position each vowel formant occupies in the range of a speaker's formants. The most important fact about the range normalization procedures to be considered here is that they require at least two points of known phonetic quality from which a speaker's formant ranges may be estimated. There are at least two such procedures discussed in the literature. The first is due to Gerstman (1968) and the second to Labonov (1971). Both procedures can be shown to reduce to the same general form, namely a linear rescaling of each formant. The general form may be stated as follows.

$$(3.1.1) \quad F^*_{N[V]s} = a_{Ns} \times F_{N[V]s} + b_{Ns}$$

The two techniques differ only in the manner in which the speaker parameters a_{Ns} and b_{Ns} are specified.[2]

Gerstman's procedure.-- Gerstman's procedure in its original form[3] may be stated as follows:

$$(3.1.2) \quad F^*_{N[V]s} = (F_{N[V]s} - F_{N[min]s}) / R_{Ns}$$

where $F_{N[min]s}$ and R_{Ns} are respectively the minimum and the range of the *N*-th formant for subject *s* over his full range of vowels. The effect of this procedure is to express the formant value of a vowel according to the position it occupies in the linear range of that formant for a given subject.

Labonov's procedure.-- Labonov's technique is viewed by that author as a means for obtaining a more reliable estimate of a

formant's position in a linear range. The procedure may be stated as follows:

$$(3.1.3) \quad F^*_{N[V]s} = (F_{N[V]s} - M_{Ns}) \; / \; S_{Ns}$$

where M_{NS} is the mean for subject *s* on the formant in question (i.e. In the notation discussed above $F_{N[\cdot]s}$) and S_{Ns} is the standard deviation for subject *s* on formant *N*.

Implications for perception.-- The details of these procedures are not important for present purposes. However, we should again emphasize that both these techniques are range normalization procedures and agree in that AT LEAST TWO VALUES with a known position on the formant range must be specified for each formant before the normalized values can be calculated.[4] A range cannot be established without at least two distinct values whose relative position in the range is known.

If human speech perception involves a process like the range normalization procedures discussed above, we would expect that at least two vowels with distinct F1 and F2 values would have to be "given" to the listener before other vowels could be categorized in relation to them.

Constant ratios and point normalization

CRH and CRH2.-- One of the earliest hypotheses for normalization is one that here will be called the constant ratio hypothesis. We will consider two slightly different but related hypotheses. The first, the constant ratio hypothesis proper or CRH, maintains that the formant values of any given subject may be derived from those of any other subject by multiplication by a SINGLE speaker-dependent constant, sometimes called a SCALE FACTOR.[5] This scale factor can in theory be estimated by the knowledge of a single formant value of a vowel of known phonetic quality. Since it requires only a single vowel point, it qualifies as a point normalization procedure.

Another related hypothesis, which we will call a two scale-factor constant ratio hypothesis or CRH2, also qualifies as a point normalization technique. It requires the specification in advance of at least one F1 value and one F2 value for vowels of known phonetic quality before other vowels can be normalized. CRH2 allows for independent scale factors in F1 and F2. The relationships assumed to hold under CRH2 are given by:

$$(3.1.4) \quad F_{1[V]s} = F_{1[V]t} \times K_{1st} \qquad \text{and}$$

$$(3.1.5) \quad F_{2[V]s} = F_{2[V]t} \times K_{2st}$$

where K_{1st} is the scale factor that relates F1 values of subject s to those of subject t and K_{2st} relates their F2 values. The single factor CRH (the constant ratio hypothesis proper) is the special case of CRH2 WITH THE RESTRICTION THAT THE TWO SCALE FACTORS BE EQUAL.

CRH is thus a stronger (more constrained) hypothesis about the nature of possible speaker variation in the formant frequencies of vowel systems. One of the most widely cited properties of CRH which does not apply to CRH2 is that under it we would expect to find F1/F2 ratios constant for a given vowel regardless of subject. As will be shown below, it is possible to transform CRH and CRH2 into additive (as opposed to multiplicative) models of normalization. Although additive and multiplicative versions are empirically equivalent, there are some conceptual advantages to the additive formulations.

Additive models of point normalization.-- A broad class of possible point normalization procedures may be represented by the following schema:

$$(3.1.6) \quad F^*_{N[V]s} = g(F_{N[V]s}) + b_{Ns}$$

where g is some transforming function applied to the formant values in hertz.

If g is chosen to be the (natural) log function, the constant ratio hypotheses CRH and CRH2 can be transformed into constant log interval hypotheses. If we use the symbol G to indicate a log transformed formant frequency, i.e., if we define

$$(3.1.7) \quad G_{N[V]s} = \log(F_{N[V]s})$$

we may then restate the constant ratio assumptions (3.1.1 and 3.1.2) as follows:

$$(3.1.8) \quad G_{1[V]s} = G_{1[V]t} + Q_{1st} \qquad \text{and}$$

$$(3.1.9) \quad G_{2[V]s} = G_{2[V]t} + Q_{2st}$$

where Q_{1st} and Q_{2st} are equal to the natural logs of K_{1st} and K_{2st} respectively (which are the F1 and F2 scale factors). These new constants will be referred to as "translation factors" since their geometrical effect on two-dimensional plots of vowel data is that of a translation in two-space.

If the translation factors G1 and G2 are constrained to be equal to each other we have the log space analogue of CRH. If there is no such constraint, we have the analogue of CRH2. The log space analogues will be referred to respectively as CLIH and CLIH2. Units in the log space will be referred to as log-hertz.

Numerical examples of point normalization

Certain relationships implicit in the constant ratio hypothesis (CRH and CRH2) are most easily illustrated by means of numerical examples using artificial "data" that exactly meet constant ratio conditions.

CRH.-- The first example to be considered is a "perfect fit" to CRH. The "data" are given in Table 3.1.1 (A). Entries for subject T are derived by multiplying the corresponding values of subject S by the single scale factor, 1.25. Thus F1 of [v] for T is 312.5, which is 1.25 times 250, the value for F1 of [v] for S.

CRH can be seen to imply that corresponding formants of corresponding vowels for different subjects show constant ratios. This is clear from the values in Table 3.1.3 (A). The ratios of F2 to F1 of each vowel are the same for both subjects: 10 to 1 for [v], 1.57:1 for [w] and 2.33:1 for [x]. Other constant ratio relationships also hold. For example, the ratio of F2 of [x] to F1 of [v] is 2.8:1 for both subjects.

CRH2.-- The "data" in Table 3.1.1 (B) represent a perfect fit to CRH2. The values for F1 are the same as in the previous example; that is, subject T's F1 values are 1.25 times the corresponding values for subject S. But the F2 scale factor is 2.0 and the F2 entries for subject T are exactly twice those for subject S. Here, CROSS-FORMANT ratios are not maintained. The F2/F1 ratio for [v] of subject S is 10:1 as before, but that of subject T is 16:1. However, ratios WITHIN FORMANTS for corresponding vowels are maintained across subjects. Thus the ratio of F2 of [x] to F2 for [v] for both subjects is .28:1.

CLIH and CLIH2.-- We will now consider the same hypothetical examples substituting values in log transformed form. The values in the lower half of Table 3.1.1 correspond to the log transformed values in the upper half.

All of the constant difference relationships observed in portion (A) of Table 3.1.1 can be seen to be represented by constant differences in portion (C). A similar remark applies to portions (B) and (D).

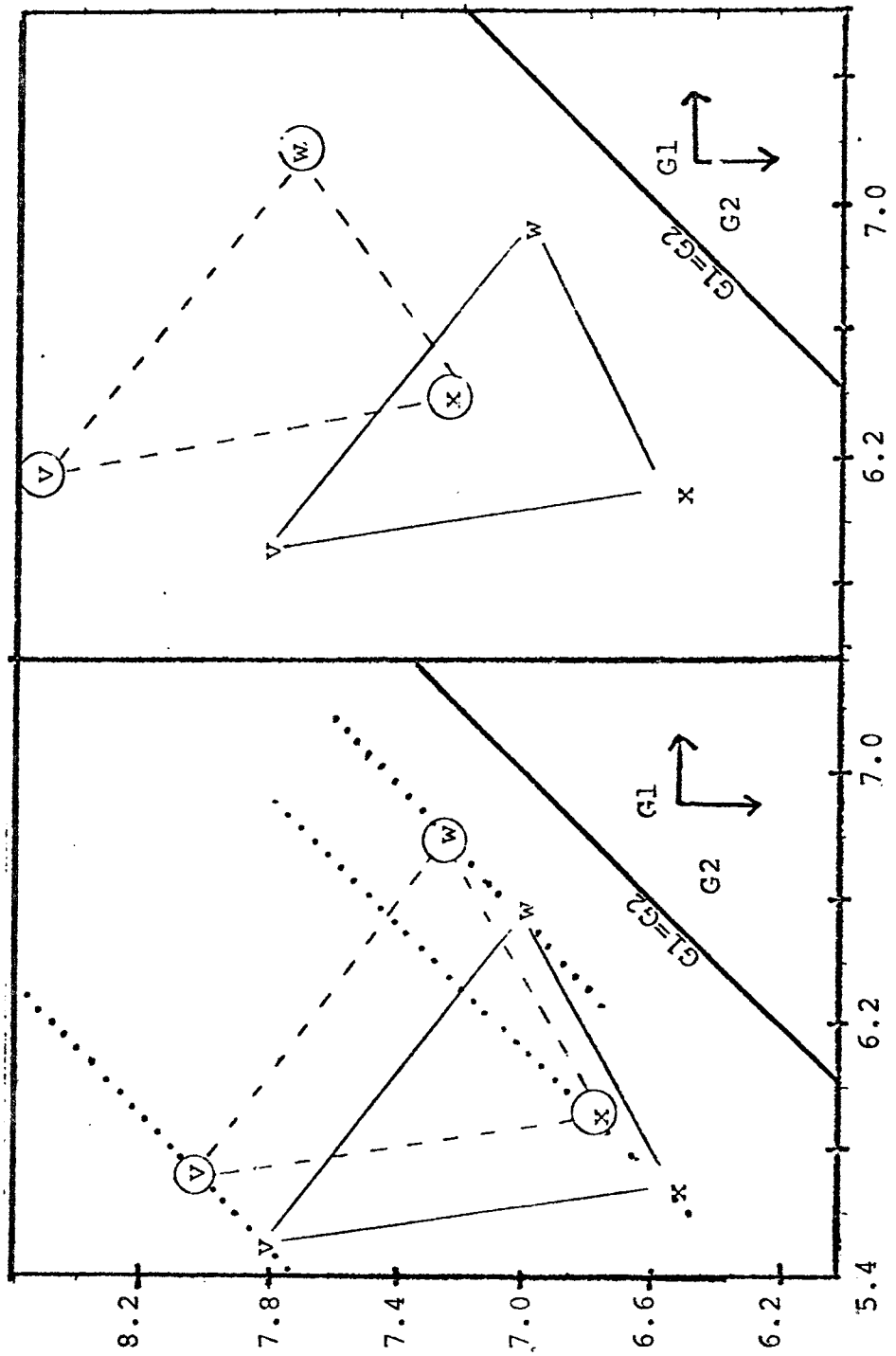


Figure 3.1.1. Graphic representation of data from Table 3.1.1.
Left half: CLIH
Right half: CLIH2

Graphic interpretation.-- A more direct indication of the relationships implied by the constant log interval hypothesis may be obtained from a graphic representation of the hypothetical data. Let us first consider CLIH2 and the data in portion (D) of Table 3.1.1. A G2 by G1 plot of this data is presented in Figure 3.1.1(B). The vowel points for subject S are connected by solid lines and those of subject T by dashed. Subject T's vowels are also circled.

The distinction between speaker-dependent and vowel-dependent information is readily discernable from this diagram. The hypothetical phonetic facts are represented by the geometrical properties of the "vowel figures" for the two subjects. The size, shape and sense (rotation with respect to the coordinate system) of the two figures are the same. In fact, a single geometrical figure can be abstracted from the diagram to represent the relationships of the vowels in the "dialect" of the hypothetical speakers.

The differences between speakers are represented by the differences in the position of the figure in the G1-G2 space. These differences in position are confined to a set of geometrical transformations referred to as translations; that is, displacements of a geometrical figure along lines parallel to the coordinate axes in which the size, shape and sense of that figure remain invariant.

The sliding template model

An intuitive understanding of the nature of point normalization and its possible use as a recognition procedure is available from a consideration of a "sliding template" model based on the geometrical properties just discussed. The vowels of the "dialect" of the speakers S and T may be represented by holes drilled in a rectangular template in a pattern that corresponds to that of the invariant geometrical figure discussed above. The template is allowed to be moved about in the G1-G2 space in the manner implied by the constraints of a translation model. It may be moved by any amount in the direction of either of the axes any number of times, but it may not be rotated.

If the [v] hole of the template is moved according to these rules to a position directly over a vowel that is KNOWN to be a given speaker's vowel [v], his other vowel points will lie under the corresponding holes of the template.

The above situation appies to the CLIH2 model, that is to the geometrical facts represented in Figure 3.1.1(B). The sliding template model may also be applied to more constrained

TABLE 3.1.1

HYPOTHETICAL DATA

VALUES IN HERTZ

(A) CRH

	Subject "S"		Subject "T"	
Vowel	F1	F2	F1	F2
[v]	250	2500	312.5	3125
[w]	700	1100	875	1375
[x]	300	700	375	875

(B) CRH2

	Subject "S"		Subject "T"	
Vowel	F1	F2	F1	F2
[v]	250	2500	312.5	5000
[w]	700	1100	875	2200
[x]	300	700	375	1400

VALUES IN LOG-HERTZ

(C) CLIH

	Subject "S"		Subject "T"	
Vowel	G1	G2	G1	G2
[v]	5.5214	7.8240	5.7447	8.0473
[w]	6.551	7.003	6.7742	7.2262
[x]	5.704	6.551	5.9269	6.7742

(D) CLIH2

	Subject "S"		Subject "T"	
Vowel	G1	G2	G1	G2
[v]	5.5214	7.8240	5.7447	8.5171
[w]	6.551	7.003	6.7742	7.6962
[x]	5.704	6.551	5.9629	7.2442

CLIH model by imposing the further restriction on the allowed movement of the template. The CLIH model requires that the template be moved equal distances in corresponding directions along both the G1 and G2 axes (again, without rotation). Translations of the template are thus restricted to motions parallel to the line G1=G2. The restrictions of CLIH implies that the vowels of the same quality will fall along constant difference lines in the G1-G2 space. The constant difference lines for the vowels [v], [w] and [x] are indicated in Figure 3.1.1(A).

A natural speech example of point normalization

While detailed consideration of the appropriateness of the different models discussed above will be postponed to Chapter IV, Figures 3.1.2 and 3.1.3 have been included to indicate the relatively good overall fit of the CLIH model to natural language data. Figure 3.1.2 represents a plot of the natural logs of the average values for the first two formants of nine American English vowels for (adult) male, female and child subjects from the study of Peterson and Barney (1952). Figure 3.1.3 represents the same data normalized according to the following algorithm:

$$(3.1.10) \quad F^*_{N[V]S} = G_{N[V]S} - G_{\cdot[\cdot]\cdot}$$

Where $G_{\cdot[\cdot]\cdot}$ is the average of G1 and G2 over all the vowels of speaker *s*. This normalization equation is in the form of the additive model given in (3.1.6) and can be derived from the relationships implicit in the CLIH model. For "perfect data", any corresponding formant value in the three vowel systems could be used as the "correction factor" with identical results. Thus, for example, we could subtract G1 of each subject's [ɛ] (instead of the mean formant value) from each of his log-formant measures. However, if we assume that formant measurements are random variables (that is that they include a random error element), the use of the mean as the correction factor will minimize the impact of random effects in individual measurements. However, it is important to emphasize that, in principle, this procedure could be applied with any single corresponding vowel known for all speakers.

Geometrical considerations of a universal figure in a formant space

In Chapter I, it was noted that a series of vowel diagrams used by traditional phoneticians until the 1920's shared several important properties with the triangle of C. F. Hellwag (1781).

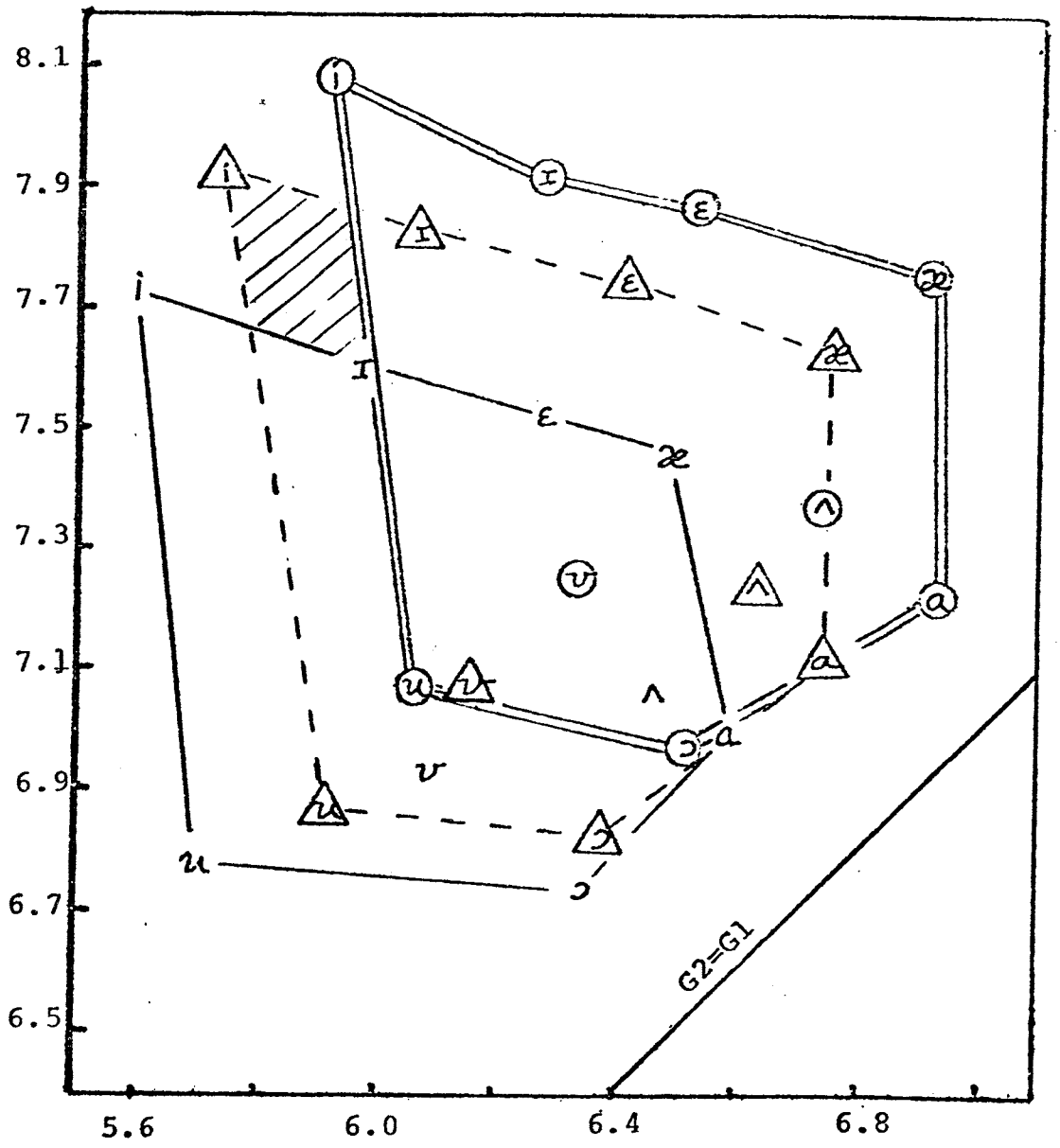


Figure 3.1.2. Unnormalized average data from Peterson and Barney (1952). Horizontal axis: G1 (=log(F1)); vertical axis: G2.

Unmarked: males

△ : females

○ : children

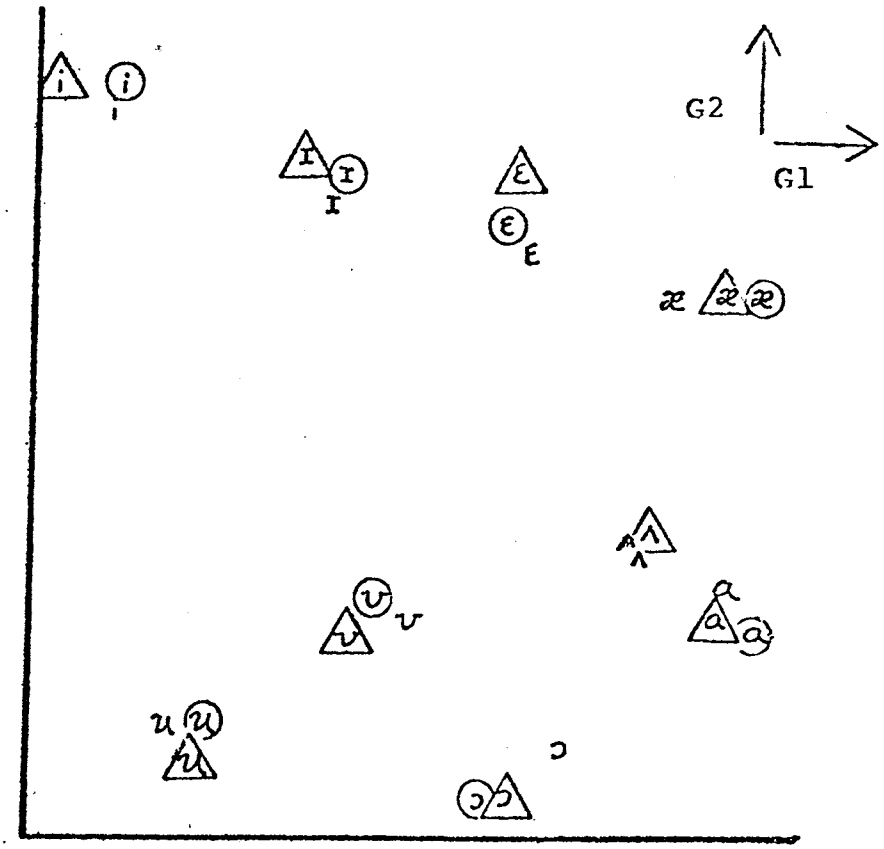


Figure 3.1.3. Average data from Peterson and Barney (1952) normalized by CLIH. Horizontal axis: normalized G1; Vertical axis: normalized G2.

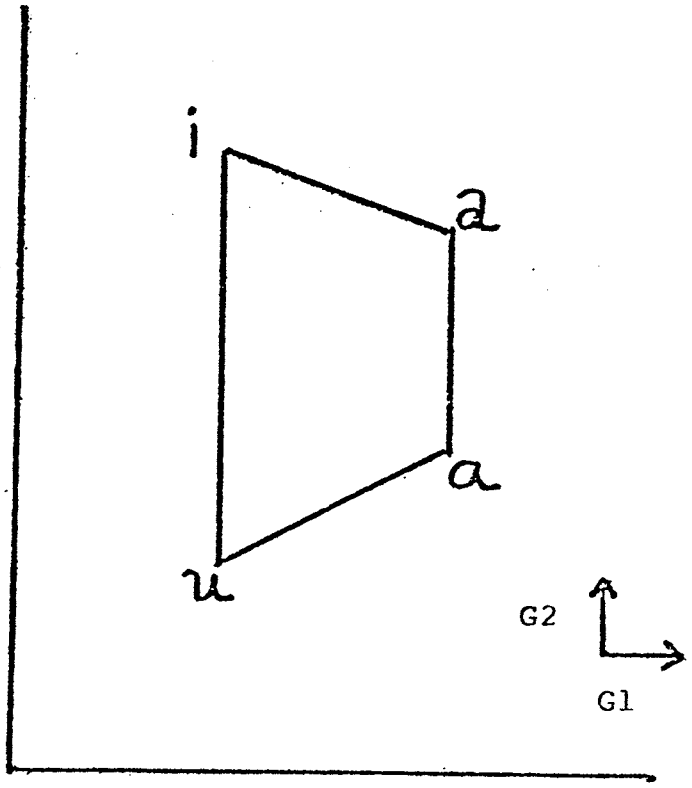


Figure 3.1.4. A Hellwagian figure in a log-formant space.

Figure 3.1.4 shows the outline of such a figure (in the appropriate rotation) placed in an arbitrary position in the G1-G2 formant space. The outline of this figure may be interpreted as the area within which the vowels of a single speaker must fall, or as the limits within which "the holes in the vowel template may be drilled." We will assume, in accord with traditional theory, that the shape of this figure is universal.

The triangle constraints.-- Once the position of the triangle for a given speaker has been determined, all his vowel formant frequencies should fall within this area. There are several interesting limitations that the universal vowel figure implies for the formant relationships of a single speaker's vowel space. These limitations will be referred to as the *triangle constraints*.

(3.1.11) No vowel in a SSS (single speaker system) has G1 lower than [i] or [u].

(3.1.12) No vowel in a SSS has a higher G2 than [i].

(3.1.13) No vowel in a SSS has a G2 lower than [u].

(3.1.14) The line [i]-[a] has a negative slope. For the front series, an increase in G1 implies a decrease in maximum G2. Points within a SSS must fall below this line.

(3.1.15) The line [u]-[ɑ] has a positive slope. For the back series, an increase in G1 implies an increase in minimum G2. Points within a SSS must fall above this line.

If a point violates the triangle constraints with respect to a certain positioning of the vowel figure, that point must lie outside the single speaker system the figure contains. If we consider Figure (3.1.2) once again, we notice that the separate outlines of the vowel areas for males, females and children bear a resemblance to the traditional vowel figure, particularly for the front vowels. The point that represents children's [æ] violates triangle constraints with respect to the vowel figure of the males' values. The experiment reported in section 3.2 below indicates that there is a tendency for listeners to fail to give high naturalness judgments to combinations of synthetic vowel stimuli near the values of males' [i] and children's [æ].

The special status of [i] under CLIH

There is another condition that is implied by the triangle constraints. It will be called the *maximum difference constraint*,

stated as follows:

(3.1.16) [i] has a greater G2-G1 difference within a single speaker system than any other vowel.

The conjunction of the maximum difference constraint with one of the implications of CLIH leads to an interesting conclusion about the vowel [i]. According to CLIH (though NOT CLIH2), the G2-G1 difference for vowels of the same quality across subjects is CONSTANT. Thus all [i]-vowels should have the same G1-G2 difference. However, the maximum difference constraint states that the G1-G2 difference is larger for [i] than for any other vowel within any given single speaker system. Thus CLIH and the maximum difference constraint imply that there exists some number, the difference between G2 and G1 of [i], that is constant across speakers and which is not shared by any other vowel. This number thus uniquely specifies the single phone [i].

Geometrically the above arguments amount to the claim that the [i]-point of a single speaker system is not contained within the area of the vowel figure of any other single speaker system. In fact, there is some support for this hypothesis in the situation represented in Figure 3.1.2. An examination of the outline of the vowel areas of the females' average data indicates that it is almost entirely included in one or both of the outlines of the males' and children's vowel areas. Only the shaded trapezoidal area near the vowel [i] does not overlap in the other two figures. There is also some perceptual evidence that the vowel [i] is well recognized in a variety of conditions (Peterson and Barney 1952, Verbrugge et al. 1976).

Normalization as a global shift in categorization functions

If the human speech perception mechanism extracts invariant relative formant relationships in a manner remotely parallel to that implied by the normalization techniques described above, we would expect to be able to induce changes in the perception of vocalic stimuli depending on the context. The expected nature of these changes is that of a GLOBAL SHIFT IN CATEGORIZATION FUNCTIONS.

The value of a categorization function for a particular speech sound at any given stimulus value may be defined as the probability of that stimulus being identified as that speech sound. If a set of categorization functions is plotted against the values along a stimulus continuum, we have categorization diagrams similar to those frequently used in the literature for summarizing experimental results. See Figure 3.1.5. In the case of a global context effect,

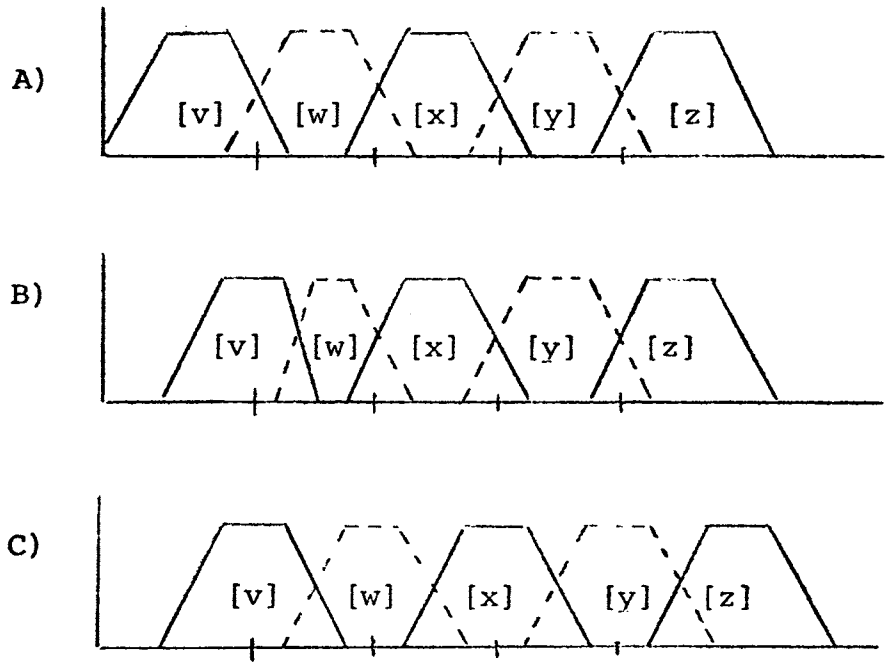


Figure 3.1.5. Categorization functions.
A) Reference functions.
B) Local shift, only [v] and [w] categories affected.
C) Global shift.

Vertical axes: percent identification
Horizontal axes: stimulus value.

we would expect a change in the categorization functions of points along the entire stimulus continuum. Such a global shift may be contrasted with a local context effect. The latter would involve a change of the shape of a categorization function near the affecting context stimulus, while leaving functions in more remote areas of the continuum unchanged. Figures 3.1.5(B) and 3.1.5(C) illustrate local versus global changes in categorization functions. According to any of the relative formant normalization schemes discussed, we would expect changes in categorization due to speaker differences to be global in scope.

Experimental evidence for the relevance of relative formant normalization to speech perception

Targets and Carriers.-- Several experiments will be described in which a distinction will be made between stimuli that are intended to provide a context of some kind and stimuli that are expected by a hypothesis under consideration to be categorized differently according to changes in the context. Following Thompson and Hollien (1970) we will refer to the former as carrier stimuli and the latter as target stimuli.

Carrier induced categorization shifts.-- Ladefoged and Broadbent (1957) present the results of the first experiments ever designed to test a general hypothesis of relative formant normalization using synthetic speech stimuli.[6] The authors analyze the effects of six different versions of a carrier sentence on the categorization of four fixed target words. The vocalic stimuli in both the carriers and targets consisted of four formants, though only the first two formants were varied in the experimental conditions (and these only in the carrier sentence). The carrier frames were versions of the sentence *Please say what this word is*, with F1 and F2 ranges raised and lowered independently from a set of neutral values. The different versions of the carrier sentence were intended to simulate speaker differences. The authors note that all but one of these versions (that with lowered F1 but raised F2) were reasonably natural sounding and "... sounded like the same sentence pronounced by people who had the same accent but differed in the personal characteristics."(1957:100)

Ladefoged and Broadbent found statistically significant evidence that the categorization of each of the four target stimuli could be changed by modifications of the formant frequencies in the carrier sentence. Furthermore, the direction of the changes was predictable in terms of the relative formant positions of the vowels within the carrier sentence with respect to the targets. There was only one case in which a predicted shift did not occur.

Evidence for normalization.-- Broadbent and Ladefoged (1960) conducted three series of experiments extending and modifying the paradigm of the experiments outlined above. In one series, the

experimenters used three test-words (target stimuli) differing from each other only in F1. As in the experiment previously described, there were several versions of each carrier sentence which had formant frequency changes corresponding to putative differences between speakers. In this case there were two carrier sentences which differed in their phonetic content; namely, versions of *This is* ---. and *What's this?*.

The results of these new experiments were also compared to a subset of the results from the experiment of Ladefoged and Broadbent (1957) involving the same test words in versions of the carrier frame *Please say what this word is*.

Both the new sentences also showed context effects. However, the different sentences produced somewhat different effects. Broadbent and Ladefoged point out one of the implications of the differences they noted in the following passage:

> ... Indeed it should be noted that the difference between *This is* and the longest sentence disproves the alternative view ... that the test word is identified as the most similar of the vowels in the introduction. *This is* contains as many samples of the vowel in *bit* as the longer sentence does, yet its effect is smaller. The other vowels in quite different parts of the spectrum must be affecting judgment. (1960:397).

That is, the effects appear to be global rather than local in nature.

Among their final conclusions, Ladefoged and Broadbent offer an interpretation of their results that has very important implications for relative formant normalization:

> Thus in our experiments, a vowel with a low formant I position will not bias subsequent perception if it is recognized as the vowel in *please* rather than in *what* -- unless the formant is placed unduly low even for a *please* (1960:398).

This conclusion may be restated as follows: the formant frequency values for a vowel are analysed by the listener into two components: speaker-dependent and vowel-dependent. Only when a formant frequency difference is analysed as a speaker difference will the categorization functions be shifted.

We may interpret this in terms of the "sliding template" model developed earlier. A change from vowel to vowel within the speech of a single speaker does not require the movement of the template and the relative positions of target stimuli within the template does not change. But when a large difference in

formant frequencies in two carriers requires that the template be moved to a new position to preserve the phonetic identity of the two carriers, the relative positions of the fixed target stimuli would be different for the two different template positions.

Perceptual evidence for point normalization

Though the distinction between range and point normalization is not made by Broadbent and Ladefoged, the results of experiments with one of the sentences gives some support to point normalization, since the conditions for range normalization are not met.

While the sentences *Please say what this word is*, and *What's this?* contain at least two distinct vowels each, and hence might serve as a basis for the estimation of a variable formant range, the sentence *This is* ---. contains only two tokens of the same vowel. That is, IT SUPPLIES ONLY A SINGLE VOWEL POINT AS A FORMANT FREQUENCY CONTEXT. Further analysis of the data of Broadbent and Ladefoged (performed by this author) indicates that differences in versions of *This is* ---. invariably produce differences (significant beyond the .01 level by a chi-square analysis) in the categorization of the three test words in the expected direction.

While this is suggestive of point normalization, it is not clear that a GLOBAL change in the categorization of vowels has been accomplished. All the test stimuli presented in combination with the *This is* ---. carriers appear to be within 150 hertz of the vowel in one of the carrier sentences with which it was presented. It is conceivable that only local effects, similar to those to be described below account for all the differences in these cases.

Local context effects for vowels

Thompson and Hollien (1970) provide evidence for what is essentially a local contrast effect for vowel targets and carriers that consist of only isolated vowels. Their experiment was designed to represent context effects that might arise within a single speaker's vowel system. In it natural and synthetic versions of the vowels [i, I, ɛ, ʌ] and [ɑ] were used as carriers while the synthetic vowels [I, ɛ] and [ʌ] along with four intermediate points, two between [I] and [ɛ] and two between [ɛ] and [ʌ] were used as targets. Their analyses indicated that there was no difference in the effects of synthetic versus natural carriers.

In discussing their results, they note:

> ... the two formant stimuli which were generated to represent the [I, ɛ, ʌ] vowels were most often identified appropriately by listeners. Moreover, responses to stimuli with intermediate formant characteristics ... indicated that they had ambiguous auditory characteristics (1970:6).

They also note:

> ... where ambiguity existed, an item tended to be assigned to a category represented by a higher F1 and lower F2 when the preceding item's F1-F2 ratio was lower than that of the item being investigated -- and vice versa ... the effect noted is one of contrast (1970:9).

However, there are indications that the effects in their experiment were primarily local in nature since the size of the context effect depended on the acoustic distance between the carrier and target vowel:

> ... the effect of an affecting carrier vowel upon the identification of the affected target vowel is seen to be greatest for samples which were closely adjacent with respect to F1-F2 ratios and tends to decrease as the affected vowel becomes less like the affecting vowel (1970:10).

Thus while the possibility of point normalization is raised by Broadbent and Ladefoged's experiment, it is necessary to test whether the alteration of the formant frequencies of a single context vowel is sufficient to induce changes in an entire set of target stimuli.

3.2 An experimental investigation of point-normalization

Introduction

The experiment to be described here was designed primarily to test the relevance of point normalization in the perception of synthetic vowels. The primary question is whether a global shift in categorization functions can be induced by the change in the formant frequencies of a SINGLE carrier vowel corresponding to the changes in natural speech resulting from a change in speaker. In order to provide a sensitive test for globality, a two-formant "continuum" of 220 target vowels in the F1-F2 space was presented in two carrier contexts. These contexts consisted of two synthetic vowels which were both highly identifiable as tokens of [i]. One of these had formant frequencies near those of average

values given by Peterson and Barney (1952) for adult male speakers and the other near those for child speakers of American English.

The results of the experiment provide strong evidence in support of a general point-normalization hypothesis. Additional evidence sheds light on the issues of a "universal vowel figure" and on the question of whether a one parameter model for speaker differences (such as CLIH) is more appropriate than a two parameter model (such as CLIH2).

Methods and materials

Both the carriers and targets consisted of two-formant synthetic vowel stimuli produced by a Glace-Holmes parallel resonance synthesizer. The parameter values output to the synthesizer were generated in real time by a FORTRAN IV program written by the author. The synthesizer output was connected through a high-fidelity amplifier to a pair of high quality headphones. Responses were entered at a keyboard and recorded on a disk for later analysis.

Carrier stimuli

The primary consideration in the construction of the carrier stimuli was that they both be highly identifiable as [i]. The first had an F1 of approximately 250 Hz, which is the same as the value given by Peterson and Barney (1952) as the average of the F1 of [i] for the 33 adult male speakers of American English included in their study. The second had an F1 of 370 Hz, which is near the average of 15 children's voices given in the same study. The F2's of these vowels were initially set near those of the Peterson and Barney data (2290 and 3200 Hz) but pilot experiments indicated that these values were somewhat too low. This may be due to the fact that the "effective second formant" of a vowel in the high front area is highly influenced by higher formants in natural speech (cf. Carlson et al. 1970). After some informal experimentation, values of 2400 and 3515 were selected as the lowest F2 values that produced reasonably good [i]'s with the two F1 values selected.

The two carrier vowels gave quite different subjective impressions as to "personal quality" of the apparent speakers both in this author's view and according to informal reports from the subjects. The vowel near the male values, whether heard in isolation or combined with a target vowel, gave the impression of an ordinary male voice (though it was unmistakably synthetic to those familiar with the synthesizer).

The second vowel was produced with the same relatively low fundamental frequency as the first (125 to 150 Hz, varying according to stress condition in carrier-target combinations). Apparently as a result of this, it was not generally heard as a child's voice, but rather as a somewhat unusual male voice of a harsh quality. It was generally agreed by subjects and casual listeners that the synthetic utterances containing the second vowel (with higher formant frequencies) gave the impression of a generally "smaller" speaker that those with the first. For this reason, combinations of targets with the carrier with the lower formant values will be referred to as the "large voice condition" and those with the higher formant carrier, as the "small voice condition".

Target stimuli

The target stimuli consisted of a grid of 220 points in an F1-F2 space. There were 15 values of F1 ranging from approximately 220 Hz to 1080 Hz in steps of approximately 55 "technical mels", a unit of measure proposed by Fant (1959) corresponding closely to the mel scale of Stevens and Volkmann (1940).[7] Twenty values of the potential 240 point grid were excluded on the criterion that F2 be at least 200 Hz higher than F1. See Figure 3.2.1.

Since the stimuli were generated in real time, it was necessary to check the calibration of the synthesizer periodically during the experiment. While no drift was detected during the course of the experiment, shortly after its completion some stability difficulties were noted for extremely high values of F1. Specifically, while the synthesizer always appeared to behave linearly up to 1000 Hz, near the highest F1 value used in this experiment, a discontinuous change of value from slightly above 1100 Hz to 1200 Hz was intermittently observed. Thus while considerable confidence can be put in the calibration as a whole, it is possible that the highest F1 value may have occasionally been higher than expected. Since no calibration errors were actually detected during the course of the experiment even for this value, it is not treated differently in any analyses.[8]

Carrier-target combinations

The stimuli were presented as pairs of di-syllables with the stress on the second syllable of each of the di-syllables. The order of the carrier and target were reversed from the first di-syllable to the second. This structure may be represented schematically as follows: [i'V](pause)[V'i], where V represents the target vowel. Both members of the di-syllables had a steady-state duration of approximately 160 msec, during which the frequency and amplitude values of the formants and the frequency of

F2 Hertz	Code															
3515	16	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
3238	15	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
2969	19	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
2711	13	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
2472	12	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
2249	11	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
2040	10	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
1846	9	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
1668	8	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
1504	7	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
1304	6	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
1191	5	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
1057	4	+	+	+	+	+	+	+	+	+	+	+	+	+	0	0
953	3	+	+	+	+	+	+	+	+	+	+	+	0	0	0	0
804	2	+	+	+	+	+	+	+	+	+	0	0	0	0	0	0
684	1	+	+	+	+	+	+	+	0	0	0	0	0	0	0	0
F1	Code	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	Hertz	21	267	317	370	419	477	534	594	654	718	785	856	927	1001	1079

Figure 3.2.1. Target stimulus grid. Points marked 0 not included.

the fundamental were stationary. The effect of stress was achieved by a difference in the fundamental of the two syllables, about 125 Hz for the first (unstressed) syllable and 150 Hz for the second. Linear transitions 120 msec in duration between steady-state values for the fundamental and formant frequencies of the syllable nuclei were supplied. Smooth formant amplitude and fundamental frequency transitions were provided leading into the steady-state of the second syllable (100 msec) and leading from the steady-state of the second (160 msec) to the surrounding silence. The formant frequency values during these periods were the same as the adjacent steady-states. A "simulated spectrogram" of a typical experimental utterance is provided in Figure 3.2.2.

Subjects' responses

Naturalness judgments.-- Subjects were asked to provide three distinct responses for each carrier-target combination they heard. Their first response was to be a "naturalness judgment" for the general appropriateness of the utterance. It was noted in pilot experiments that certain combinations of carrier and targets were markedly less natural sounding than others. The conditions under which this "unnaturalness" arise may be related to the "triangle constraints" discussed in the preceding section.

There were four categories of naturalness judgment, labelled "OK", "?", "BAD", and "VERY BAD".[9]

Categorization of carriers and targets.-- The last two responses recorded were categorizations of the carrier vowel and target vowel respectively. For both, subjects were given the option of 12 vowel category responses: There was one additional response labelled "other" for vowels that did not seem to the listeners to fit into any of the other categories. One subject, a native speaker of German, remarked that he used the "other" category for vowels in the [ø]-[œ] area. The other subjects used this category only rarely.

Subjects were instructed to "try to hear the carrier vowel as [i]", but that if the carrier clearly sounded like some other vowel, to respond with that vowel instead.[10]

Subjects were instructed to categorize the target vowels by comparison to the vowels of their own dialects of American English.

Phonetic backgrounds of the subjects

The subjects were all graduate students in linguistics at the University of Connecticut. All but one of the five were

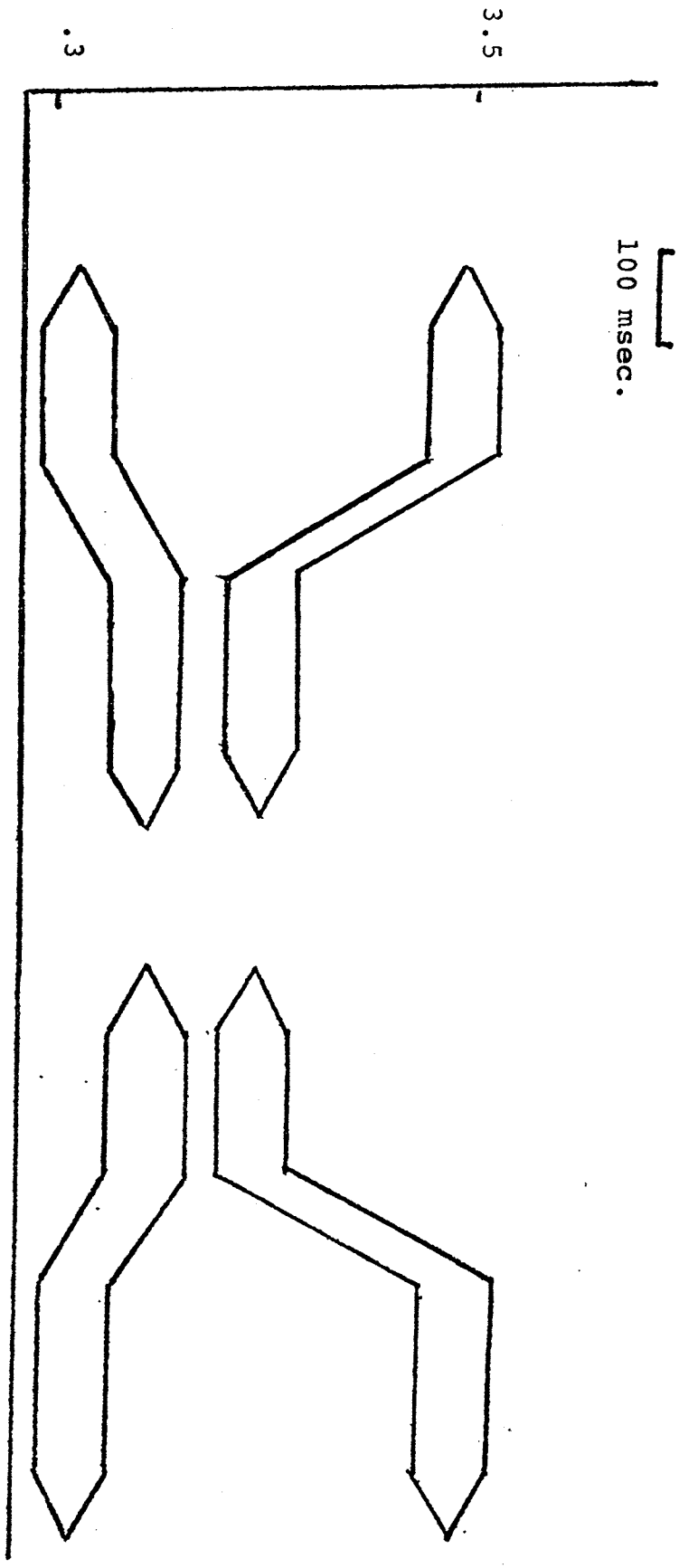


Figure 3.2.2 Simulated spectrogram of typical target-carrier combination. Vertical axis in kilohertz.

native speakers of American English. The remaining subject, a native speaker of German, speaks American English with native fluency and for all practical purposes may be considered a native speaker of American English.

All of the subjects had had prior experience in phonetic transcription. However, as noted above, subjects were asked to use vowels of their own dialects as standards for categorization. There are indications of definite differences in the subjects' categorizations of the target stimuli which may be due to dialect differences. Nevertheless, data from the five subjects has been pooled for most analyses since the increased sample size more than compensates for the variability in listener-dependent categorization.

Experimental sessions

Each experimental session consisted of the presentation of the entire 220 vowel continuum in one of the two voice conditions, that is either with the large voice or small voice [i] as carrier. There were four separate randomizations of the target stimuli, each presented once to each subject in both voice conditions. The voice conditions and randomizations were themselves randomized across subjects with the restriction that voice conditions alternate on successive sessions.

Because of the amount and complexity of the information demanded of the subjects in their responses, the presentation program was designed so that subjects could listen to a particular utterance as many times as they pleased before entering either the naturalness or the phonetic category judgments. They always heard each utterance at least twice: once before the naturalness judgment was entered and once between the naturalness judgment and the phonetic categorization of the carrier and target. As a consequence, the length of the experimental sessions varied somewhat from subject to subject, and from session to session. The average duration of a single session was approximately one hour.

The entire experiment represents about 40 subject-hours of data, eight hours each by five subjects. A total of 5280 responses (1760 each for naturalness, carrier and target judgments) was recorded from each subject.

Subjects were paid at the rate of three dollars per hour.

The primary appropriateness conditions

There are two cases in which a combination of carrier plus target will be referred to as relatively "inappropriate". The

first of these involves what will be termed "carrier shift", or conditions in which the carrier vowel is categorized as a vowel other than [i]. Within a general point normalization framework, we are testing the hypothesis that the target vowels are being categorized as a function of the relationships of their formant frequencies to those of a carrier vowel of a FIXED phonetic quality. If the phonetic quality of the carrier vowel changes, the interval relationships it bears to the target points are changed.

The problem may be visualized in terms of the sliding template model developed in section 3.1 above. Assume a vowel template of fixed size, shape and sense is placed on the formant space such that the carrier vowel appears under the template hole for [i]. The target vowels are then presumably categorized according to their proximity to holes in the template as it is currently positioned. If the carrier vowel were interpreted as a vowel other than [i] (say as [e]), a different part of the template will have been placed over the carrier stimulus, and hence the other holes in the template will be in positions different from when the carrier stimulus was categorized as [i].

The second case of inappropriateness involves the stimuli that were categorized as relatively unnatural. Since most stimuli were categorized in the relatively higher naturalness categories, there is reason to withold the more unnatural vowels from the initial analysis.[11]

We will therefore consider a subset of the responses which we will refer to as the *primary (appropriateness) conditions*. Responses meeting the primary conditions are those which were given either of the two highest naturalness judgments in which the carrier was ALSO categorized as [i]. Though the primary conditions constitute only a fourth of the eight possible naturalness-carrier combinations, they constitute nearly half of the responses for both voice conditions (49.5 percent for the large voice and 45.5 per cent for the small).

Predominance boundary plot

The most general way of dealing with categorization in the two formant space is to consider the results in terms of a set of three-dimensional categorization functions for each vowel. The X and Y axes are respectively the F1 and F2 values of a stimulus and the Z axis is the number (or percentage) of responses for the particular vowel category in question. A full set of three-dimensional categorization functions cannot readily be plotted. However, a graphic technique which we will call a "predominance boundary plot" has been devised to represent differences in the categorization in a two formant space. This

procedure attempts to delineate areas in the F1-F2 target space within which a particular categorization (in the primary appropriateness conditions) predominates over all other responses.

Figure 3.2.3 represents predominance boundary plots for the combined data of the five subjects. Boundaries for vowels in the large voice condition are delineated with solid lines. Those for the small voice condition are marked with dashed lines. The areas for each vowel response are identified by the phonetic symbols along the boundary curves. Those for the small voice are enclosed in circles. The positions of these vowel labels indicate local boundary points as calculated by an algorithm outlined in Appendix I. The lines are hand drawn so as to delineate continuous areas of predominance.

It is clear from this plot that a general upward shift in the categorization functions of all the vowels is represented in this plot. Only the F1 boundary of [æ] does not show a clear upward movement. Evidence for the globality of the shifts is strongest in the case of vowels in the [ʌ]-[ɑ]-[ɔ] area since these are phonetically most remote from the changes in the carrier stimuli. In short, there is strong indication that categorization functions are shifted in roughly the manner predicted by a point normalization hypothesis.[12]

While this plot provides rather striking evidence for the globality and systematic nature of the categorization shifts, other methods are required for a more detailed and quantitative measure of the strength of the effects.

Partial categorization functions and t-tests

Partial categorization functions.-- Separate two-dimensional "partial categorization functions" for F1 and F2 may be plotted. These correspond to the marginal distributions of the three-dimensional functions described above. Thus for F1, the value of such a function for a given vowel at each F1 stimulus level may be obtained by summing the number of responses given to that vowel across all F2 levels.

Plots of such partial categorization functions for F1 are provided in Figure 3.2.4 for the front vowels and in Figure 3.2.5 for the back. While the marginal distributions are less complete than the full categorization functions, they have the advantage of being generally easier to deal with statistically.

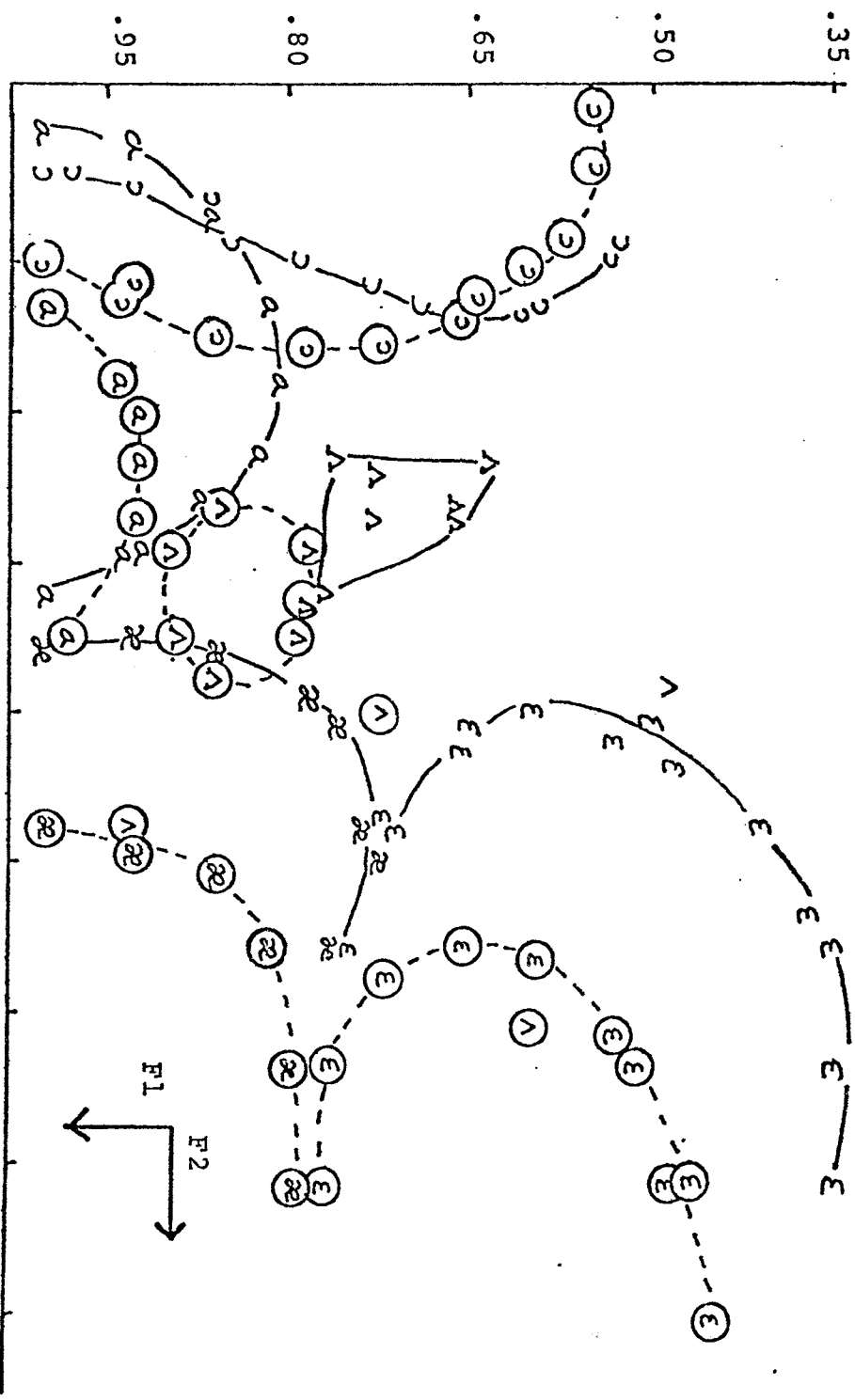


Figure 3.2.3. Predominance boundary plot. Axes in kilohertz. See text.

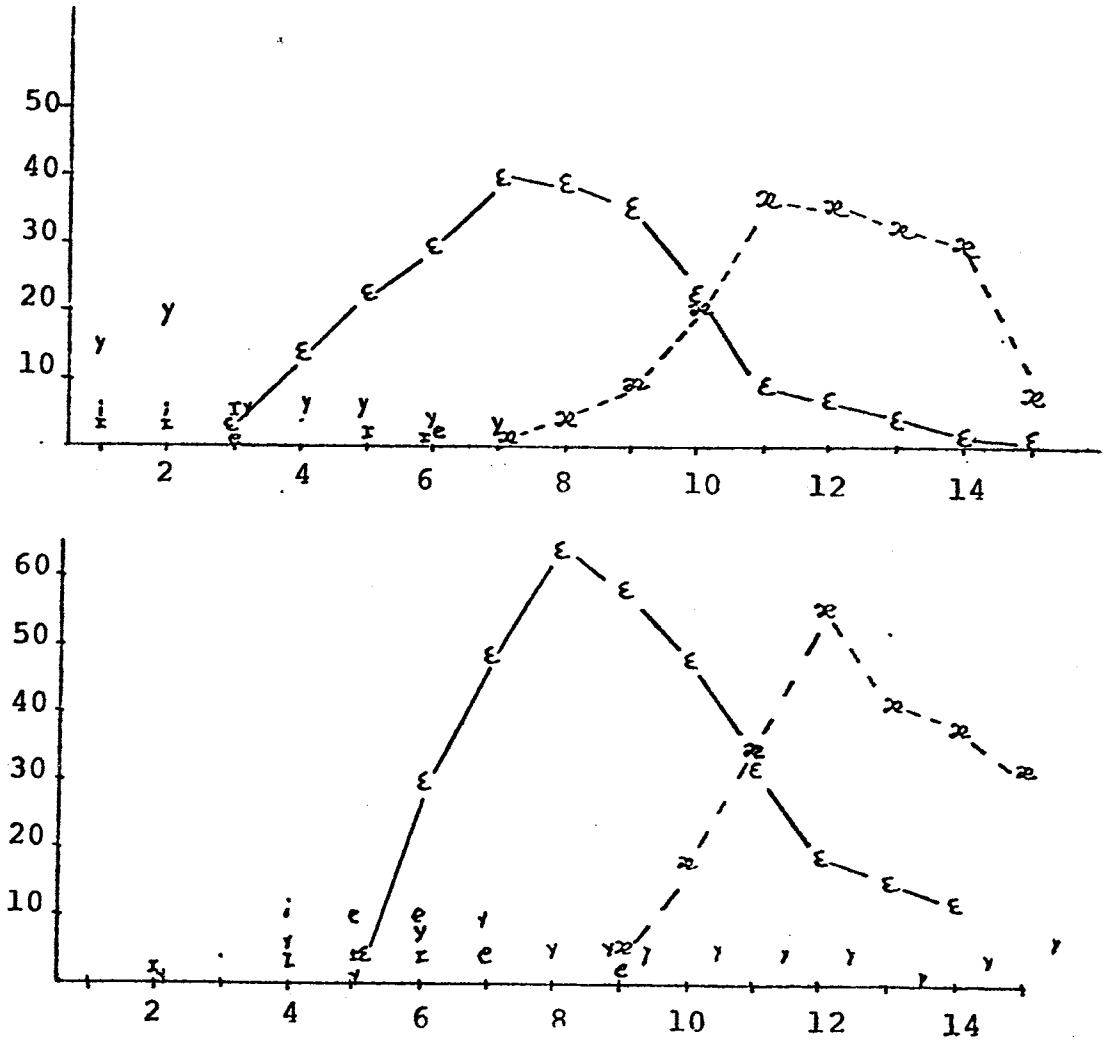


Figure 3.2.4. Partial categorization functions for F1 of front vowels. Upper half: large voice. Lower half: small voice. Horizontal axes in code values of F1. Vertical axes: number of primary condition responses.

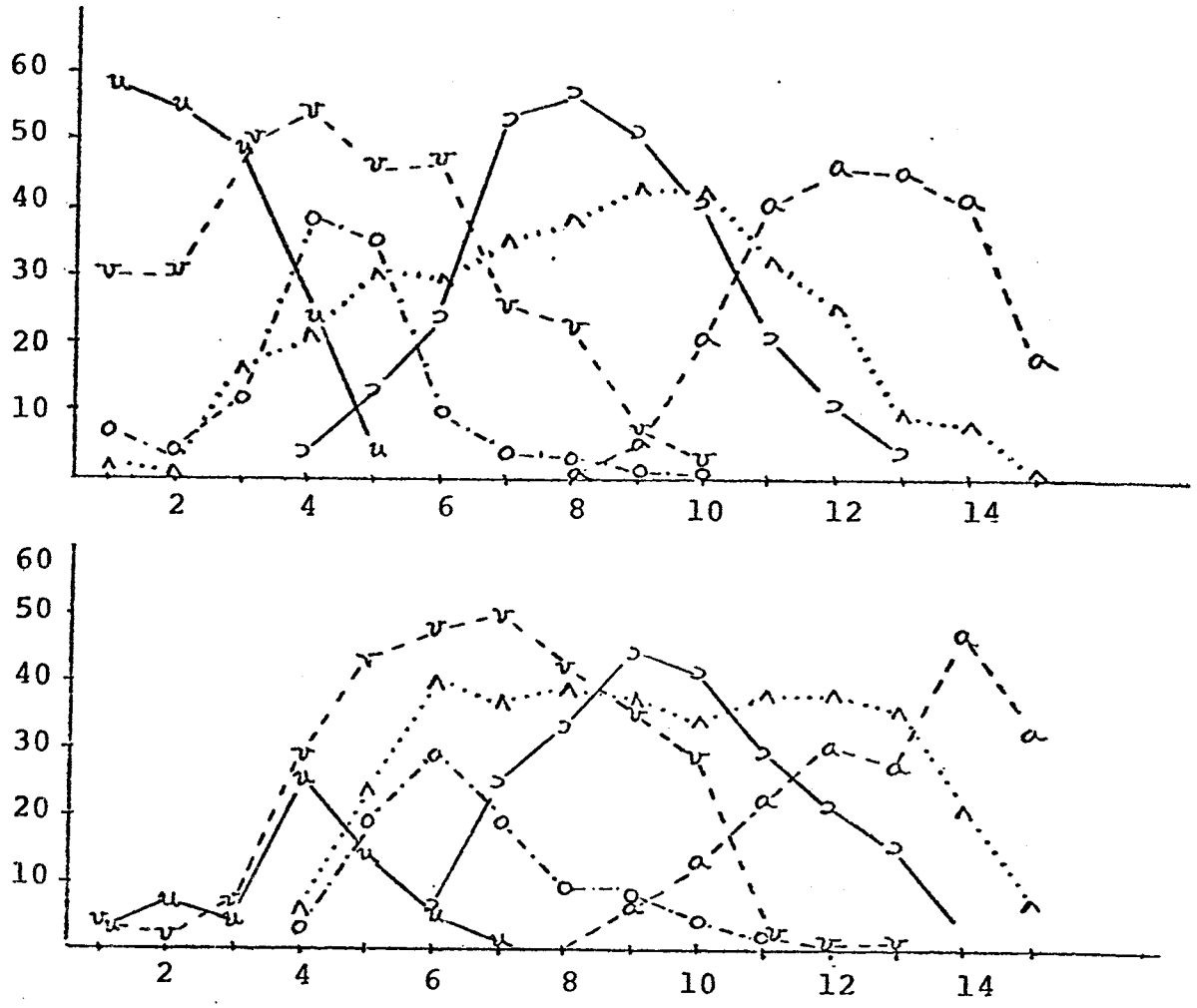


Figure 3.2.5. Partial categorization functions for F1 of back vowels. Upper half: large voice. Lower half: small voice. Horizontal axes in code values of F1. Vertical axes: number of primary condition responses.

T-tests.-- Assuming the responses to any category are roughly normally distributed about the mean, we may use the t-test for the equality of the means of categorization functions for a given vowel in each of the two voice conditions.[13]

A series of t-tests using the actual Hertz values rather than the code values of the stimuli[14] were performed to test for differences in the categorization of each vowel in the two conditions. Results for F1 and F2 are presented in Table 3.2.1.

The mean value of the stimuli categorized as in the two conditions has shifted upwards in the predicted direction for all the categories of each of the vowels. All of the differences are significant beyond the .05 level in a one-tailed test.

Comparison with overall categorization.-- The results for categorization in the primary conditions may be compared to those for the stimuli overall (without regard to naturalness or carrier combination). All of the categorization changes in the overall responses are again in the expected direction, though that of F1 of [æ] is only marginally so. See Table 3.2.2.

This comparison indicates that the differences associated with the two voice conditions are generally larger in the "more appropriate" stimuli. In F1, 34 of 45 available comparisons show that the difference between the two voices is larger for the primary conditions. The primary conditions show larger differences in F2 in 36 of 45 cases.

The triangle constraints and appropriateness

Carrier shifts.-- There is some information relevant to a point normalization hypothesis in the patterning of carrier-target combinations that result in carrier shifts (responses to the carrier other than [i]). These patterns are related to some of the constraints on combinations implied by a universal (Hellwagian) vowel figure.

The general shape of a Hellwagian figure indicates that there should be no vowels in the system with F1 lower than [i] and none with F2 higher. (See triangle constraints 3.1.11 and 3.1.12.) In a substantial number of cases for both voice conditions, one or both of the implied constraints would be violated if the carrier were interpreted as [i]. It is in such situations that most carrier shifting occurs.

Figure 3.2.6 shows the number of carrier shifts at each target stimulus value for the large voice carrier. Most of the stimuli with F2 higher than the carrier cause shift.

TABLE 3.2.1

CATEGORIZATION OF TARGETS IN PRIMARY CONDITIONS

Vowel	Mean Small Voice Hz	Mean Large Voice Hz	Diff.	t	df
F1					
i	396	273	122	4.5665	24
I	335	330	105	4.0271	32
e	490	419	71	2.7695	43
ɛ	664	578	86	7.2393	564
æ	899	855	44	3.7248	402
ʌ	693	614	79	5.8126	695
a	923	875	48	4.0193	520
ɔ	697	620	77	7.1577	499
o	517	424	93	7.4309	239
U	526	400	126	13.1396	614
u	368	282	86	10.2411	269
y	651	314	337	10.0563	124
F2					
i	3032	2301	731	2.9500	24
I	2811	2173	638	6.5819	32
e	3248	2564	684	6.8579	43
ɛ	2879	2311	568	16.1005	564
æ	2664	2066	598	16.5089	402
ʌ	1884	1605	279	9.9283	695
a	1638	1148	220	7.1983	520
ɔ	1092	998	94	6.1612	499
o	977	888	89	4.0264	239
U	1808	1447	361	10.4163	614
u	1157	1055	102	2.1047	269
y	2080	1831	249	2.1330	124

TABLE 3.2.2

OVERALL CATEGORIZATION OF TARGETS IN HERTZ

Vowel	Mean Small Voice	Mean Large Voice	Diff.	t	df
			F1		
i	328	269	59	3.8700	213
I	371	285	86	7.0327	135
e	472	402	69	4.9184	264
ɛ	635	583	52	5.7943	1275
æ	905	904	1	.1245	1190
ʌ	697	644	53	4.1753	879
a	961	928	33	3.8780	773
ɔ	711	636	75	7.7474	754
o	505	434	71	7.3839	426
U	468	391	78	8.6310	950
u	307	277	30	6.4707	919
y	366	294	72	3.9952	514
			F2		
i	3351	2963	388	4.8722	213
I	2922	2435	487	6.9312	135
e	3189	3038	151	2.3038	264
ɛ	2923	2750	173	6.1616	1275
æ	2876	2639	237	7.8842	1190
ʌ	1892	1631	261	10.1555	874
a	1663	1485	178	5.9811	793
ɔ	1084	987	97	7.0293	754
o	932	871	61	3.2965	426
U	1844	1507	337	9.4011	950
u	1431	1133	298	7.2287	919
y	2289	1866	423	6.1330	519

18	19	18	18	19	18	18	17	17	14	16	18	17	17	17
19	18	19	18	19	20	17	18	18	18	16	18	16	18	18
19	20	18	19	19	18	17	19	19	19	16	16	18	17	17
19	19	19	17	20	19	16	17	17	16	16	16	19	17	16
11	13	15	15	13	17	17	15	17	16	15	18	16	18	18
5	3	3	4	2	0	0	2	2	2	3	3	3	7	8
2	1	3	3	2	1	0	2	1	2	4	1	3	3	3
4	3	2	1	1	3	1	0	1	1	0	2	3	3	2
4	2	1	0	1	1	1	0	0	1	2	1	1	3	3
2	2	0	0	2	0	0	0	1	4	0	1	2	2	4
1	2	1	2	1	2	1	1	0	0	2	2	2	2	2
0	1	0	2	0	1	1	1	0	1	1	0	0	0	2
2	0	0	0	1	1	1	0	0	1	0	1	2		
1	2	3	0	3	0	0	3	2	1	2				
0	1	0	1	2	1	1	2	2						
1	2	1	2	0	0	3								

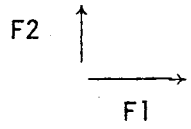


Figure 3.2.6. Carrier shift for large voice. Axes in stimulus code values. Symbol (x) marks approximate carrier stimulus position.

20	18	20	6	(x) 6	5	5	3	4	6	7	2	5	9
19	19	19	12	5	3	1	2	5	1	4	1	2	3
20	18	20	9	6	1	1	0	0	0	0	3	0	2
19	19	18	9	5	1	0	3	0	0	0	1	2	0
19	18	17	10	4	1	1	1	0	0	0	0	0	0
20	20	19	14	5	0	1	0	0	1	0	0	1	1
20	18	18	15	0	1	1	1	1	0	0	1	1	1
18	20	17	16	5	2	1	0	2	0	1	1	0	2
18	19	20	11	9	2	0	0	1	1	0	0	0	0
18	18	18	11	10	7	0	0	3	1	1	0	1	2
16	19	20	10	5	1	1	2	1	1	1	0	2	0
18	17	16	16	8	5	1	5	0	0	1	1	1	1
18	18	18	10	9	1	0	3	0	2	0			
17	19	19	11	3	0	0	0	1	2				
19	18	18	12	9	7	2	1	0					
19	16	17	16	8	2	2							

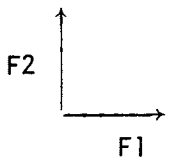


Figure 3.2.7. Carrier shift for small voice. Axes in code values. Symbol (x) marks approximate carrier stimulus position.

Figure 3.2.7 indicates that for the small voice, carrier shift occurs for most target stimuli that have F1 lower than the carrier. The nature of the shifts in both cases is generally what would be expected by a relative formant hypothesis. Nearly all cases of shift involving a target with a too high F2 show the carrier categorized as [y].[15] When the target F1 was too low, the carrier was categorized as [e] and more rarely as [I].

According to triangle constraint 3.1.14, not only are there no vowels with F2 higher than [i] but F2's of vowels with higher F1's should have lower F2's than [i]. An analogous argument applies to the back vowels: we expect that the vowels with higher F1's than [u] will also have higher F2's. In the case of the back vowels, there is a limit on the stimulus space, since the target vowels are constrained such that there are no stimuli below the line F1=F2+200Hz. There is no such definitional constraint on the stimuli at high F2 levels; there the stimulus space is completely rectangular and all F1's occur with all F2's. The triangle constraints lead us to expect that combinations of a carrier categorized as [i] and a target of high F1 and F2's the same or even somewhat lower than the target should be so incompatible. There is limited support for this in the large voice vowels in the carrier shift pattern. (Figure 3.2.6.) Thus at F2 level 11 which is slightly below the F2 of the carrier, the highest two F1 stimuli result in a total of 15 carrier shifts, while the lower 10 F1 levels produce a total of only 22. There is no evidence of a similar pattern in the small voice condition. However, evidence for the sloped side aspects of the triangle constraints may be found in the pattern of "most natural" target-carrier combinations.

Naturalness judgments.-- Figure 3.2.8 represents the total number of highest naturalness judgments given to combinations of the large voice carrier and each of the targets (regardless of carrier shift and vowel categorization). Inside the heavy dashed lines in this diagram, the utterances have received at least 5 (out of a possible 20) naturalness judgments at the highest level. The heavy lines thus bound the most peripheral stimuli which received at least five judgments at the highest naturalness level. The corresponding information for the small voice condition is presented in Figure 3.2.9.

The general shape of the areas contained manifests some of the properties of a Hellwagian figure. The pattern is not entirely clear since there appears to be a strong reduction in the high naturalness judgments given to stimuli in the area of the carrier stimulus for both voice conditions and in a region of low F1's for

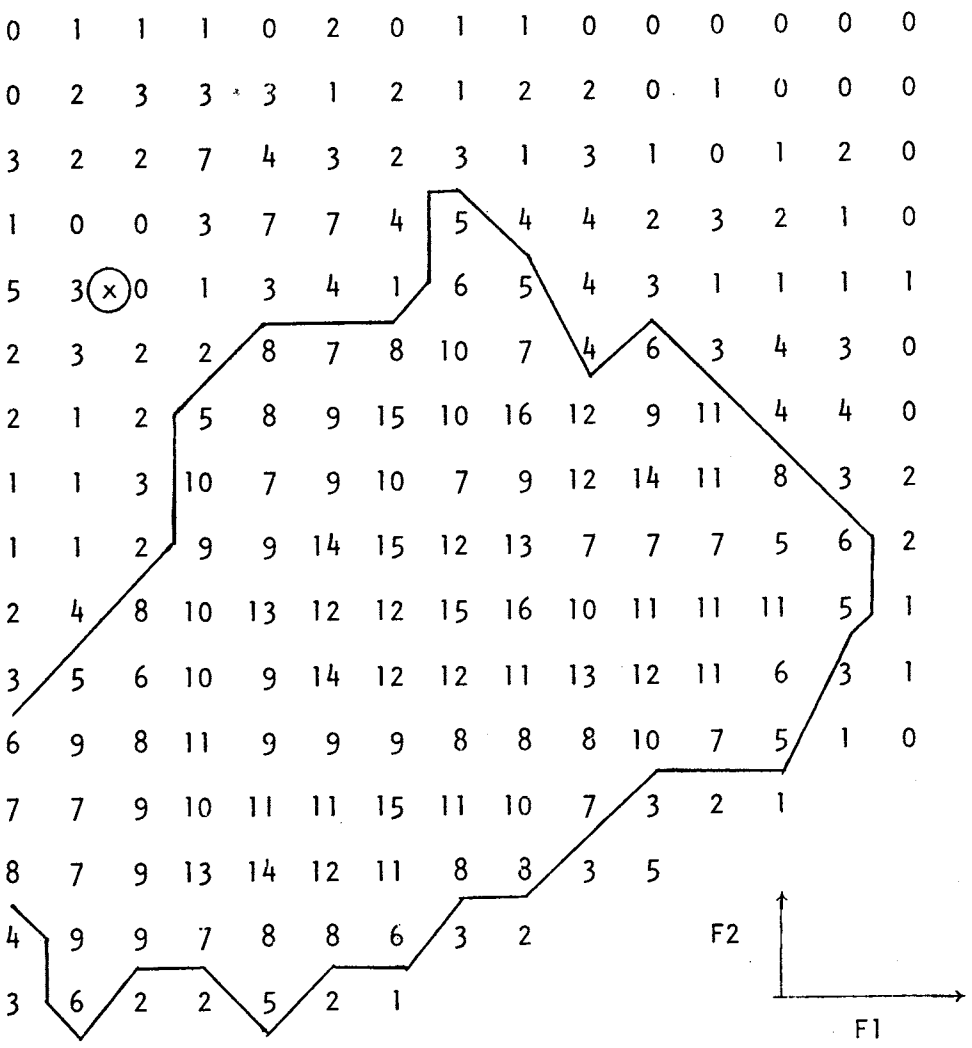


Figure 3.2.8. "Best" naturalness judgments for large voice. Axes in code values. Symbol (x) marks approximate carrier stimulus position.

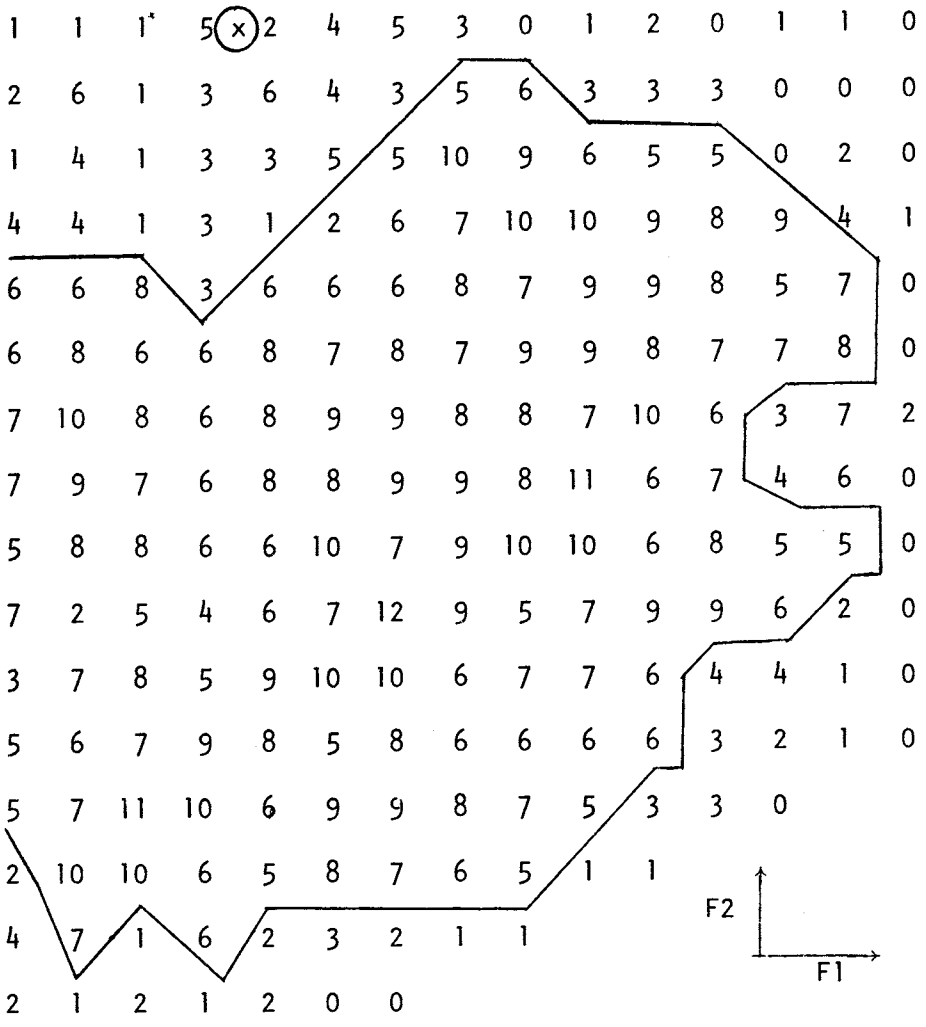


Figure 3.2.9. "Best" naturalness judgments for small voice. Axes in code values. Symbol (x) marks approximate position of carrier stimulus.

the large voice. However, there is evidence for a "roughly triangular" shape at higher F1 levels for both the front and back vowels. This pattern is generally shifted upward in F1 and F2 with the change from the large to the small voice carrier.[16] While the evidence is at best suggestive, one might interpret the heavy lines as "outlines of the vowel area of the template" when the [i] holes are in place over the respective carrier stimuli. Further research concentrating on stimuli near the periphery of the putative vowel area would be necessary to clarify this issue.

Relationship between carrier shift and naturalness.-- As pointed out in section 2.3.1, given the maximum difference constraint (3.1.16) and CLIH, we are forced to the conclusion that the vowel [i] is uniquely specified by its formant frequencies alone, without regard to any other vowels of a single speaker system. If these conditions were strictly true of human speech, we would expect that a "true" [i] would never be categorized as any other vowel. In the present experiment, we would then expect that carrier vowels would not shift. The results noted above show that carriers do in fact shift when the triangle constraints are violated.

However, lower naturalness judgments are relatively more frequent for the carrier shift than when the carrier is heard as [i]. There is thus an additional penalty in appropriateness of combination when carrier vowels are forced to shift. A chi-square contingency table analysis indicates that the association between carrier shift and naturalness is highly significant for both voices (large voice chi-square = 679.18, df = 3, $p < .001$; small voice chi-square = 94.17, df = 3, $p < .001$; see Table 3.2.3).[17]

These associations provide weak evidence supporting a model like CLIH, which (together with the maximum difference constraint) implies a "privileged position" for the vowel [i]. Under a two parameter point normalization model, such as CLIH2, there would be no reason to expect that carrier shift should be in any way inappropriate.

Quantitative comparison to constant ratio hypotheses

It is possible to compare the predictions of the constant ratio hypothesis to the results of the present experiment. To do this, we compute for each vowel a quantity to be referred to as a "change ratio ". The change ratio for a given formant of a given vowel is defined as the ratio of the mean of the categorization function for the small voice condition to that of the large voice condition. According to CRH, we would expect the

TABLE 3.2.3

CARRIER SHIFT BY NATURALNESS

	"V. BAD"	"BAD"	"?"	"OK"
		LARGE VOICE		
NON-SHIFT	169	535	1116	1043
SHIFT	412	476	485	148
NON-SHIFT	276	739	1149	840
SHIFT	226	286	499	317

values of all the change ratios in both F1 and F2 to be the same. However, the carrier stimuli themselves (which were adjusted empirically without regard to formant ratios) do not exactly meet the constant ratio hypothesis since the change ratio in F1 of the carriers is about 1.35 while that in F2 is 1.45. The overall change ratios for all stimuli meeting the primary conditions were about 1.15 for F1 and 1.20 for F2. However, the constraints on the target space make it seem unwise to regard the actual quantitative results as definitive. The fact that the target space is bounded causes difficulties for vowel categories that occur near its periphery. Some aspects of this problem are considered below.

Change ratios in F1.-- In order to investigate the constancy of change ratios within a formant, we may plot the change ratio for a given vowel on that formant against the average of the means of the categorization functions. Consider the case in which the mean for the target stimuli categorized as [i] in the small voice condition is 375 Hz and the mean for the large voice is 275 Hz. The change ratio for these vowels is 375/275. This will be the Y-coordinate of the [i] point. The X-coordinate is the mean of the means of the large and small voice /i/'s or (375+275)/2 = 325 Hz. We will refer to a plot of such points as a change ratio plot.

On such a plot, if the change ratio were constant (as predicted by CRH2 for a single formant) we would expect to find change ratio values distributed along a line parallel to the X-axis. On the other hand, local contrast effects such as those observed by Thompson and Hollien (1970) would demonstrate a pattern in which the size of the change ratio decreased abruptly with increasing distance on the X-axis for the carrier stimuli (which are located at low F1 values). Figure 3.2.10 shows such a change ratio plot for F1 using the means from Table 3.2.1 (uncircled points). The change ratio does not remain constant, but appears to be decreasing somewhat as a function of the level of F1. The predictions of CRH2 are not strictly born out by this data.

Further analysis indicates that there are "mitigating circumstances" in this apparent lack of fit. First, we may question the validity of the very extreme change ratio for the vowel [i], since this vowel is generally underrepresented in the data (as are [e] and [ɪ]). Only two of the five subjects actually show any [i] target responses in the primary conditions for both voices. There is also evidence for a larger shift at higher F1 values than is indicated by means of [æ] and [ɑ]. The distribution of the responses for these two vowels may be running against the limits of the synthesis space. There is some

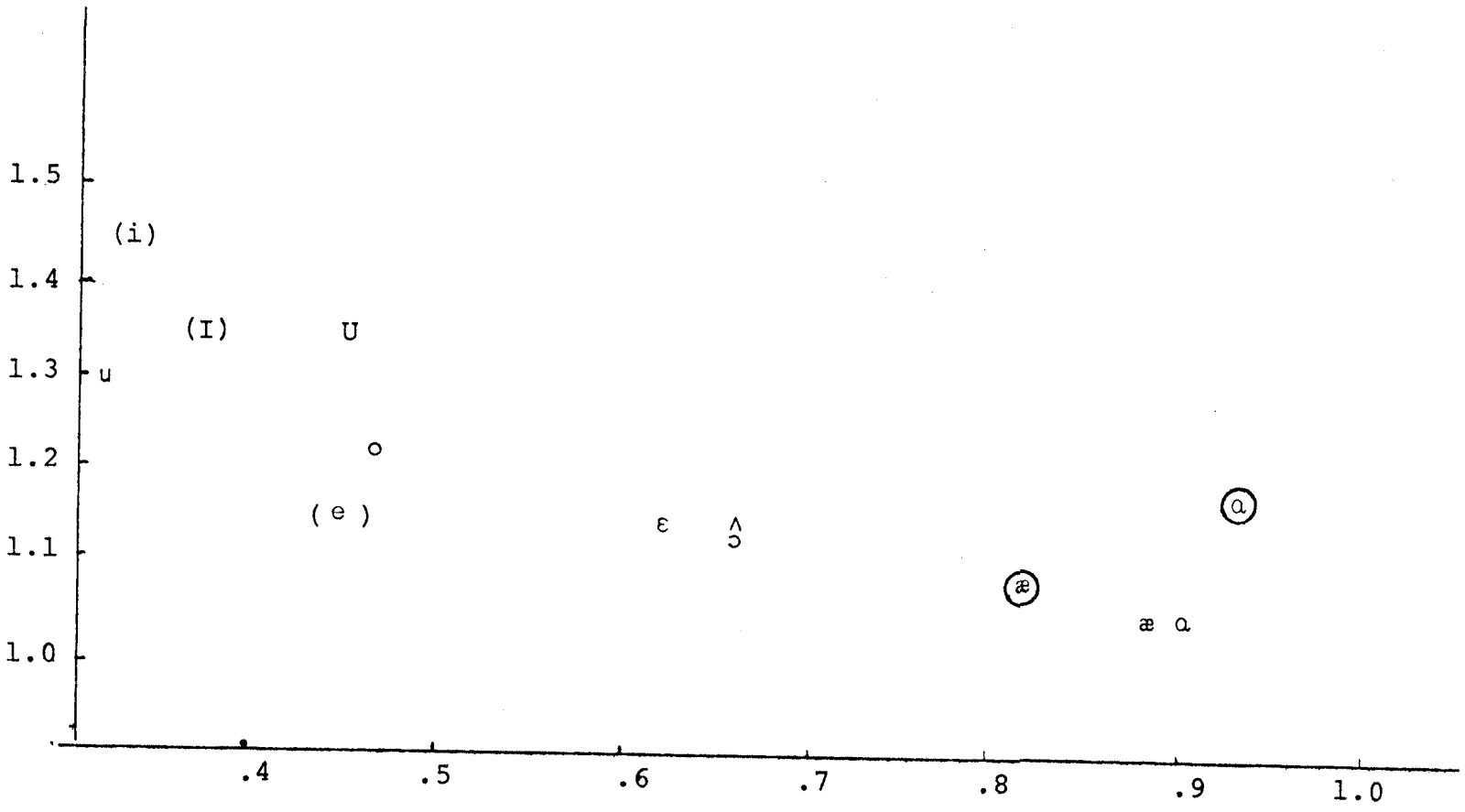


Figure 3.2.10. Change ratio plot for F1 of "primary condition" data. Vertical axis is change ratio. Horizontal axis is "average of means" of categories in kilohertz. See text for details.

indication of this in the plot of the marginal F1 distributions given earlier (Figures 3.2.4 and 3.2.5). The mode of the partial categorization function for [æ] is a stimulus value higher for the small voice than for the large. This corresponds to a difference of 71 Hz. The number of categorizations for the small voice is still quite high even at F1 level 15 while it has dropped off considerably for the large voice. (This is a result of vowels at the highest F1 level being categorized outside the primary appropriateness conditions.)

The shift in the modes for [ɑ] is also larger than indicated by the means. Here the mode of the categorization function has moved up two stimulus levels, or about 143 Hz. Figure 3.2.10 includes change ratios for [æ] and [ɑ] calculated on the modes rather than the means of the partial categorization functions. These points are enclosed in circles. When these points are considered, it appears that while there are still somewhat larger change ratios at low F1 values, especially for [u] and [U], they remain fairly constant above 500 hertz. Such a pattern might tentatively be interpreted to indicate a local contrast effect SUPERIMPOSED on a global shift in categorization functions.

Change ratios for F2.-- A change ratio plot for F2 is presented in Figure 3.2.11. As in the case of F1, the general pattern does not show a constant change ratio, but rather one which decreases with distance from the F2 of the carriers in question (the carrier vowels have high F2 values). It is again possible that this pattern represents a combination of local contrast and global shift.

In this case, however, a consideration of female-male data from the study of Peterson and Barney (1952) indicates that natural change ratios may not be constant but rather decrease with the frequency level of the vowel formants involved.[18] A change ratio plot for the Peterson and Barney data is presented in Figure 3.2.12. It shows a generally similar pattern to the change ratios of the means of F2 values found in this experiment (Figure 3.2.11). The analagous comparison of F1 change ratios in this experiment with the Peterson and Barney data does not show such a resemblance.

Further observations

Effective second formant.-- While the mean F1's for vowels categorized are generally in the range of those found for corresponding natural data, the F2 values for non-back vowels are sometimes considerably higher than natural F2's. However,

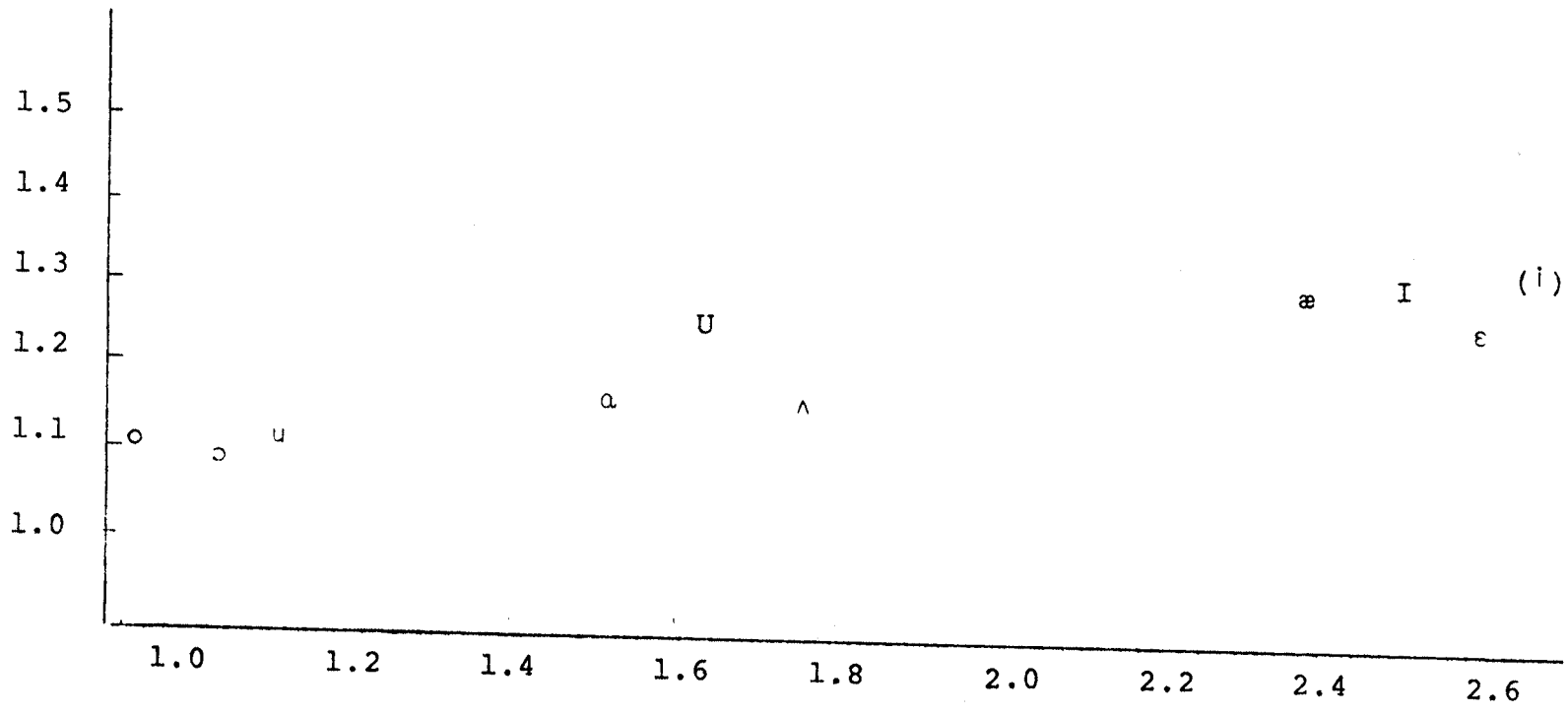


Figure 3.2.11. Change ratio plot for F2 of "primary condition" data. Horizontal axes in kilohertz.

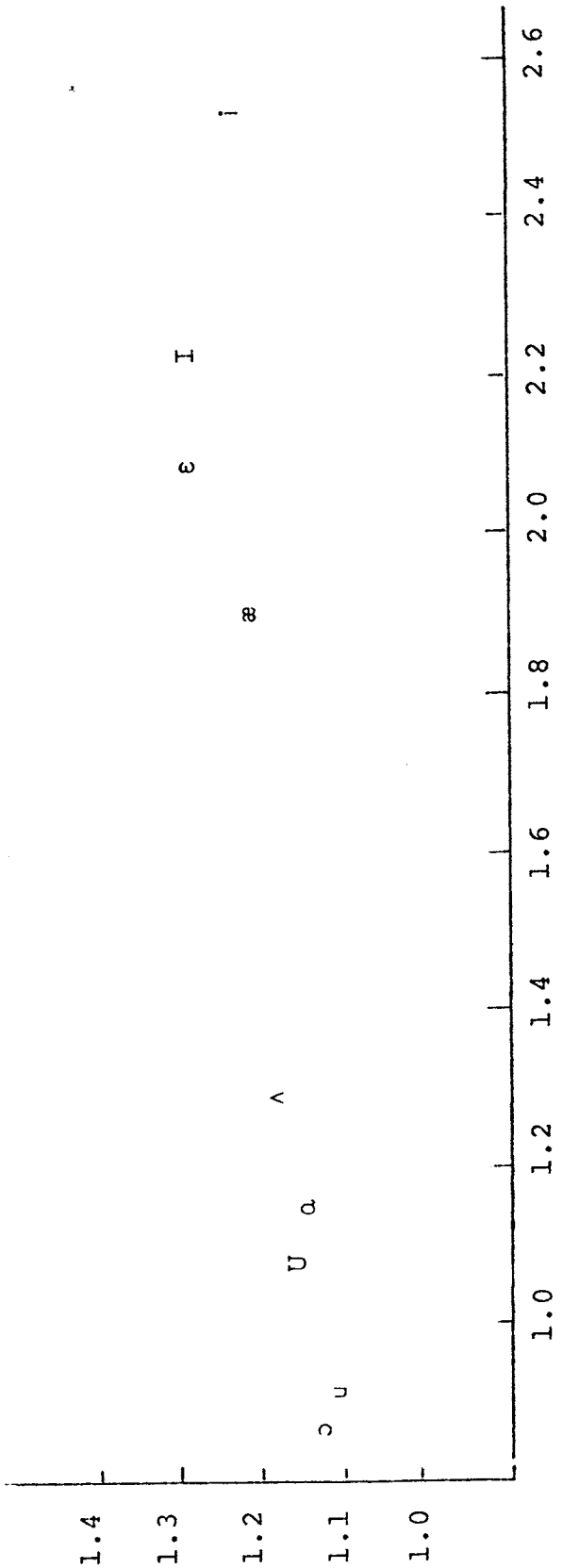


Figure 3.2.12. Change ratio plot for F2 of Peterson and Barney (1952) data. Horizontal axis in kilohertz.

the F2 values in this experiment are generally comparable to the results of the early categorization experiments of isolated synthetic vowels presented by Miller (1953).

The fact that two formant synthetic vowels seem to require higher F2 values than corresponding multiformant natural vowels was noted early by Delattre et al (1951). The results of the present experiment corroborate this result. This may indicate that an effective second formant (F2', cf. Fant 1959, Carlson et al 1970) which depends to some degree on energy in higher formants should be considered in the analysis of natural data.

It has occasionally been suggested that the effective second formant might be more nearly constant for different speakers than F2 alone. If the means of the F2's of this experiment were representative of F2' values, the present experiment argues against such a claim, since mean F2 values actually showed a greater average ratio than mean F1 values.

Persistence of shift effects over time.-- In addition to the results discussed earlier, Broadbent and Ladefoged (1960) found a marked tendency for a reduction in the magnitude of context effects induced by carrier sentences on target vowels with prolonged exposure to the carrier stimuli. Such a result would be generally contrary to any of the relative formant normalization hypotheses considered here. However, it is possible that this "wearing off" of context effects is attributable to the small number of target stimuli used in that experiment. Such a small set may allow them to serve as their own reference system, allowing subjects to ignore the carrier sentence. In the present experiment, subjects are exposed to each of 220 target stimuli only once in each session and there is no possibility for the targets to serve as a reference system.

As a rough test of the effects of prolonged exposure to the experimental situation, two series of comparisons were made with the stimuli that met the primary appropriateness conditions. The first series compared the differences in the means of the categorizations in the two voice conditions using only the first 50 stimuli from each session. The second series used only the last 50 in each session.[19]

There is no indication of a difference between the two conditions of exposure. Within the first 50 stimuli, all 22 of the formant comparisons (11 each from F1 and F2) show a positive shift from the large to the small voice condition. In the last

50, 21 of 22 show the predicted effect.

Comparisons of the size of the differences (rounded to the nearest tenth of a code value) in the two exposure conditions indicated that of 18 available comparisons, 10 were greater in the later stimuli, 8 less and 2 were the same. Thus, the magnitude of voice effects appears to be little changed by increased exposure to the carrier stimulus.

Conclusions

Although the quantitative effects of shifts in categorization functions are not exactly what would be expected according to either CRH or CRH2, the fact that the change of a single reference point produces a global shift in categorization functions provides strong support for a general point normalization hypothesis.

Perhaps the fact that the carrier vowels were forced to occur with targets in combinations that violate the triangle constraints has interfered with the quantitative results of this experiment. Future experiments might be so designed that both carriers and targets were varied in different voice conditions. The present experiment was designed to provide minimal differences in the carrier context. Further experiments specifically designed to discriminate between normalization models, providing alternative carrier contexts, would seem to be in order.

NOTES TO CHAPTER III

[1]It has been suggested that F3 and possibly higher formants combine with F2 to produce an "effective second formant", or F2'. F0 and occasionally F3 have been suggested as factors in the possible solution of the vocal tract normalization problem described below (cf. Miller 1953, Peterson 1951 and Nordström and Lindblom forthcoming).

[2]We can represent the class of relative formant normalization procedures to be discussed here by the following schema:

$$F^*_{N[V]S} = f(F_{N[V]S}, P_S)$$

where f is some function and P_S is a vector (an ordered list) containing speaker-dependent parameters for subject *s*. The form of the function and the nature of the contents of the speaker information vector are specified by the particular normalization procedure in question. Such a specification formally meets the requirements ofour stated goal, the establishment of a functional relationship between physical parameters and putative phonetic features. The question of the empirical adequacy of "features" thus extracted, of course, remains.

[3]Actually, what is described below is only a part of Gerstman's procedure. In the original article Gerstman (1968) also uses rescaled sums and differences of F1 and F2, as well as rescaled F3.

[4]This is reflected formally by the fact that the normalization function contains two speaker-dependent parameters.

[5]Though methods of estimation of this scale factor have varied, CRH has been suggested (with varying degrees of explicitness) by Chiba and Kajiyama (1941), Peterson (1951, 1961), and by Nordström and Lindblom (forthcoming), among others.

[6]All the results to be discussed involve synthetic speech targets. The importance of relative formant information in the phonological identification of vowels in natural speech appears to be less than that of other factors. See Verbrugge et al (1976). It should be noted that experiments involving identification of natural speech involve the identification of phonological elements. In English, vowel quality differences constitute only a part of the phonetic contrasts separating phonological vowel categories.

[7]The formula for the calculation of technical mel values from hertz measurements may be given as:

$$m = 1000 \log (x/1000 + 1) / \log(2)$$

where x is the frequency value in hertz. The mel scale was chosen on the basis of pilot experiments that indicated that linear divisions in hertz provided too many stimuli at higher values of F1 and F2 so that high vowels were over represented in categorizations along the F1 axis and front vowels along F2.

[8]Carrier-target combinations involving the highest F1 value do show a marked tendency to be judged as relatively unnatural by the subjects, especially for the large voice condition.

[9]Subjects were instructed to use "OK" as the usual response and to use the other judgments for utterances they thought sounded worse than average. Though two of the five subjects appeared to have behaved in more or less this way, two others tended to give middle values. One of the subjects showed a very high rate of "BAD" and "VERY BAD" responses.

[10]This option was provided on the basis of remarks made by listeners during the course of informal pilot studies during which subjects spontaneously remarked that the carrier vowel did not always sound like [i].

[11]Possible reasons for the unnaturalness of certain stimuli from the point of view of a point normalization model are discussed in a separate section below.

[12]Plots for individual subjects were also prepared. All subjects showed patterns in which most of the boundaries were shifted upwards. The smaller numbers of data points produce less well defined boundaries. Individual differences in categorization were apparent, particularly for the vowels [U] and [ʌ]. However, the boundaries for the individual subjects, though displaced in position, usually showed the same shapes in the two voice conditions.

[13]Since this condition is not exactly met, the probabilities associated with the differences can only be regarded as approximate. Although it is customary in experimental designs to perform an analysis of variance for vowel effects as a whole before proceding with individual t-tests, this was not done in the present case. Since ALL the changes in the means of the categorization functions were in the predicted direction, a sign test on the differences is sufficient to indicate the significance of the overall effects ($p < 5 \times 10^{-8}$).

[14]Similar tests run on the code values resulted in very nearly the same t-values in all cases. The hertz values seemed generally more informative since the size of the differences are more directly interpretable.

[15]Even subjects who showed very few [y] and [e] responses for target vowels gave them to carriers in these conditions.

[16]The reasons for the decreased levels of naturalness in areas near the carrier is not clear. One possibility is that the transitions between the carrier and target are too long in duration for such a small change in the formant frequencies.

[17]The overall naturalness judgments are not significantly different for the two voices, by a contingency table analysis. The large voice shows a somewhat larger proportion in carrier shift (34.7%) than the small (31.4%). Though this difference is not great, it is significant beyond the .01 level. The main point here is that the experimental design was reasonably successful in providing a target space that was about equally compatible with both carrier vowels.

[18]This has been suggested by Fant (1966, 1975) and will be explored in some depth in Chapter IV. Female-male change ratios have been chosen for comparison with the results of the present experiment because the average change ratios are nearer to those of the experiment than are the child-male ratios.

[19]Since each stimulus occurred only once in each session, we cannot directly compare the results of the last 50 with the first 50 stimuli within a voice condition. However, since the same four randomizations were used in both voice conditions, exactly the same target stimuli are involved when we compare the first or last 50 stimuli in the two voices for all sessions.

CHAPTER IV

RELATIVE FORMANT NORMALIZATION HYPOTHESES AND NATURAL DATA

Introduction

In the last chapter, the conceptual groundwork was provided for several relative formant normalization techniques. It was shown that perceptual experiments indicated the psychological relevance of point normalization: the specification of a single vowel is sufficient to cause a shift in an entire set of categorization functions. Though the quantitative results were not conclusive, the changes were not radically different from predictions of CRH.

In the present chapter, detailed analyses of natural formant data (i.e. measurements of formant frequencies of vowels in natural speech) are presented. In section 4.1, a large sample comparison of four normalization procedures is presented. It is shown that CLIH and CLIH2 (the log-additive versions of CRH and CRH2, respectively) compare well with the less constrained range normalization hypotheses (though Lobanov's procedure provides slightly higher identification).

In section 4.2, analyses of variance based on the CLIH2 model for a large sample are presented. The analyses indicate that additive speaker and vowel dependent effects account for about 91% of the total variation in G1 measurements and about 95% in G2.

While the results of the analyses of sections 4.1 and 4.2 indicate a generally excelllent fit for a log-additive hypothesis, section 4.3 explores certain objections raised by Fant (1966, 1975) against the notion of constant ratios of vowel formants among male, female and child speakers. While Fant's (1966) earlier hypotheses concerning the physiological origins of such discrepancies are not born out by recent computer modelling (Nordström 1975), suggestions of systematic departures from constant ratios persist.

In section 4.4, one aspect of the suspected departures from constant ratios -- the correlation of scale factor values with formant frequencies -- is subjected to further analysis. It is suggested that an additive hypothesis based on a transformation somewhat different from log may lead to a better account of speaker differences in vowel formants.

The gains over the simple log-additive hypothesis appear marginal, at best. However, this chapter is intended primarily as an illustration of certain techniques that may be applied to a traditional problem of experimental phonetics. Perhaps the most substantial empirical contribution to be made here is an indication of the generally good fit of relative formant hypotheses to natural data.

4.1 A Large Sample Comparison of Four Normalization Procedures

The procedures to be compared

The four normalization procedures chosen for evaluation are those described in section 3.1 above: the two point normalization procedures, CLIH and CLIH2; and the two range normalization procedures proposed by Gerstman (1968) and Lobanov (1971). The same subscript notation used in Chapter III will be adopted here except that an additional subscript for replications will be added, since two tokens of each vowel from each subject are included in the data to be analysed. Thus $F_{N[V]sr}$ is the measurement (in hertz) of the Nth formant of the rth token of vowel v for subject s. $G_{N[V]sr}$ is the (natural) logarithm of $F_{N[V]sr}$.

The four algorithms in the computational forms used in the following analyses are stated below:

CLIH: $$F^*_{N[V]sr} = G_{N[V]sr} - G_{\cdot[\cdot]s\cdot}$$

CLIH2: $$F^*_{N[V]sr} = G_{N[V]sr} - G_{N[\cdot]s\cdot}$$

Gerstman: $$F^*_{N[V]sr} = (F_{N[V]sr} - F_{N[min]s\cdot})/ R_{Ns}$$

where $F_{N[min]s\cdot}$ is the minimum of $F_{N[V]s\cdot}$ over all the vowels of subject s on formant N.

Lobanov: $$F^*_{N[V]sr} = (F_{N[V]sr} - F_{N[\cdot]s\cdot})/ S_{Ns}$$

where S_{Ns} is the standard deviation of formant N for subject s.

Data

The data to be used in this analysis consists of the INDIVIDUAL SPEAKERS' formant measurements on which the Peterson

and Barney (1952) study was based. Measurements of F0, F1, F2 and F3 for TWO tokens of each of the 10 American English vowels [i, I, ɛ, æ, ɑ, ɔ, U, ʌ, u, ɝ], from each of 76 speakers are provided. The vowels were spoken in [h_d] frames. The speakers consisted of 33 (adult) males, 28 (adult) females and 15 children. Peterson and Barney (1952) are careful to point out that the dialect backgrounds of the speakers are quite mixed. This should be born in mind in the subsequent analyses. As in the previous chapters, we will confine our analysis to the frequencies of the first two formants. We will also exclude from consideration the vowel [ɝ], since this vowel is generally considered to differ from the others along a separate phonetic dimension, known as "retroflexion" or "r-coloring".

Further details of the data collection and measurement techniques are available in Peterson and Barney (1952). Henceforth this data will be referred to simply as the P&B data.

Measures of resolving power

The degree of success of a normalization procedure depends in large measure on its resolving power, that is on its ability to allow the separation of normalized formant values into distinct groups corresponding to phonetic categories. While the notion of resolving power seems intuitively clear, the question of how to compare different algorithms on such a basis is not. Several possible measures of resolving power are presented below.

Graphic classification.-- The first proposed measure is the percent identification of vowels on the basis of manually drawn boundaries in two-dimensional plots of normalized formant frequencies. The graphs used for this analysis are presented in Figures 4.1.1 to 4.1.4. Only misidentified points are labeled. The classification lines are also provided. Such a technique has the advantage of directly representing the results of normalization graphically. It has the disadvantage of subjectivity in the drawing of the boundaries.

One- and two-dimensional coefficients of resolution.-- A second measure of resolving power to be proposed consists of two numerical indices which will be referred to as *coefficients of resolution*. While we are generally interested in the resolution of the data points in a two-formant space, we will also be concerned with the one-formant resolution of vowels along G1 or G2 axes.

As a natural candidate for a coefficient of resolution in the univariate (single formant) case, eta-square (η^2) is proposed. This is the square of Fischer's correlation ratio and corresponds to THE PROPORTION OF THE TOTAL VARIANCE IN NORMALIZED FORMANT

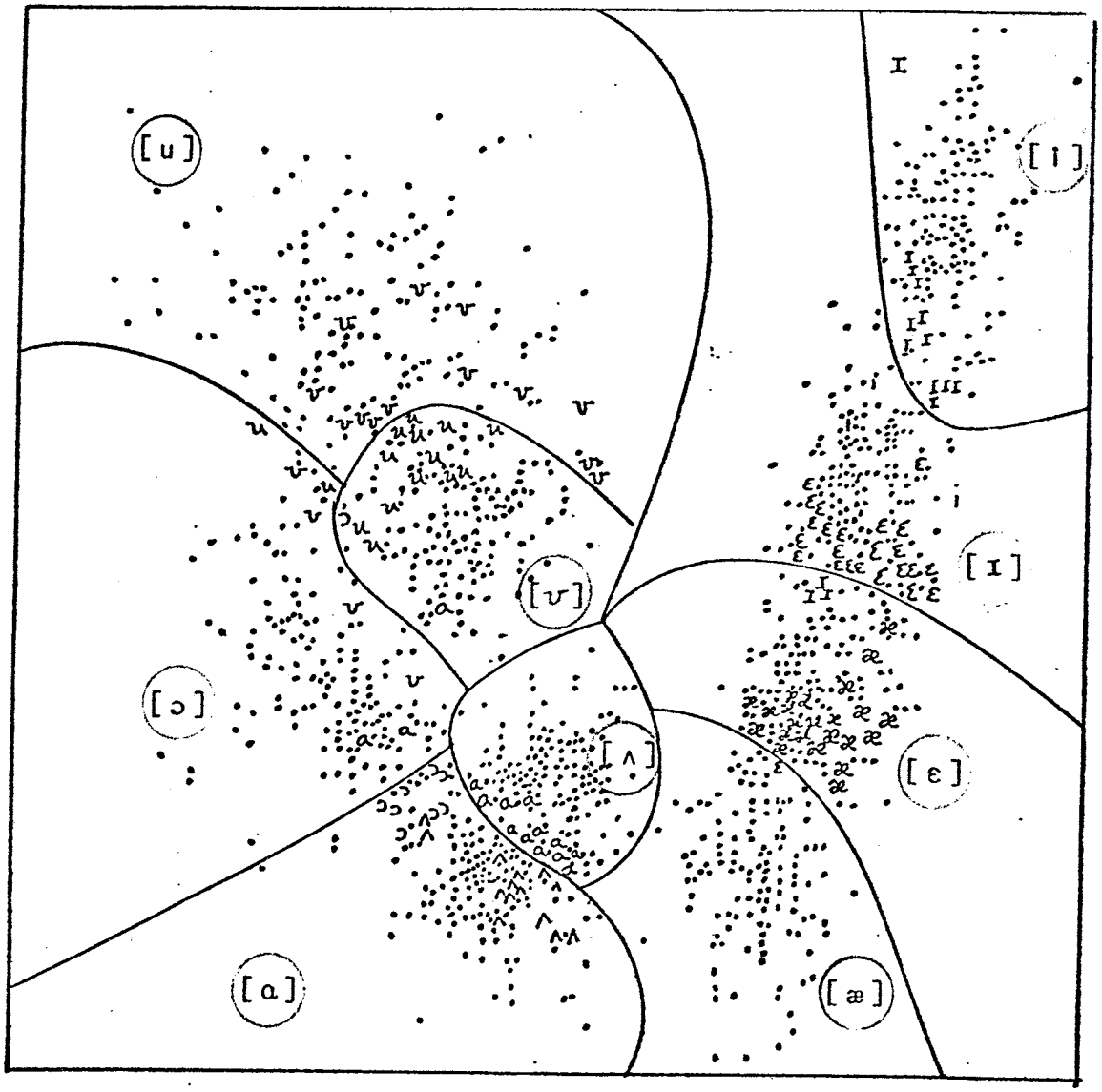


Figure 4.1.1. CLIH normalization of P&B data. Correctly identified points are indicated with dots.

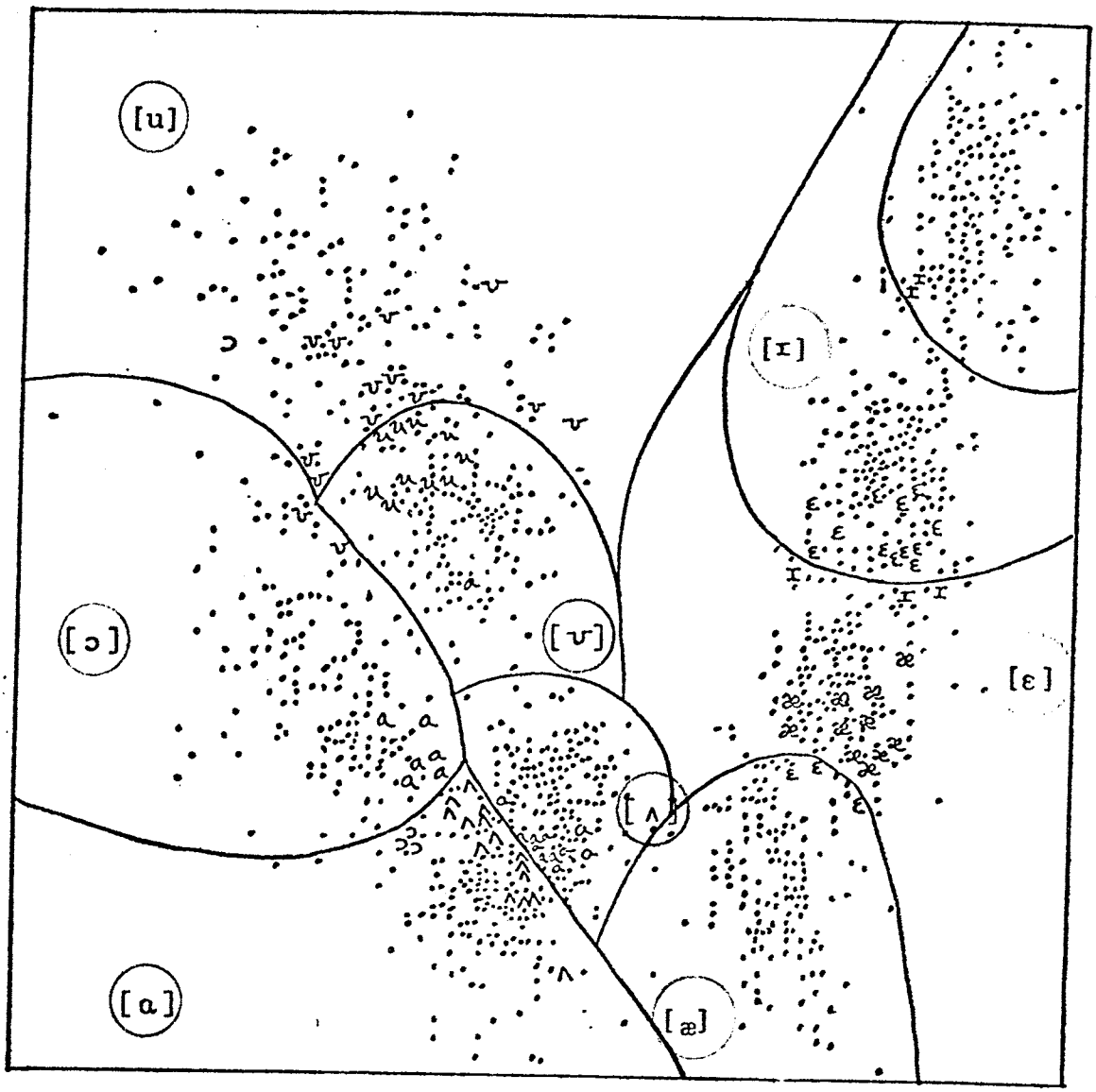


Figure 4.1.2. CLIH2 normalization of P&B data.

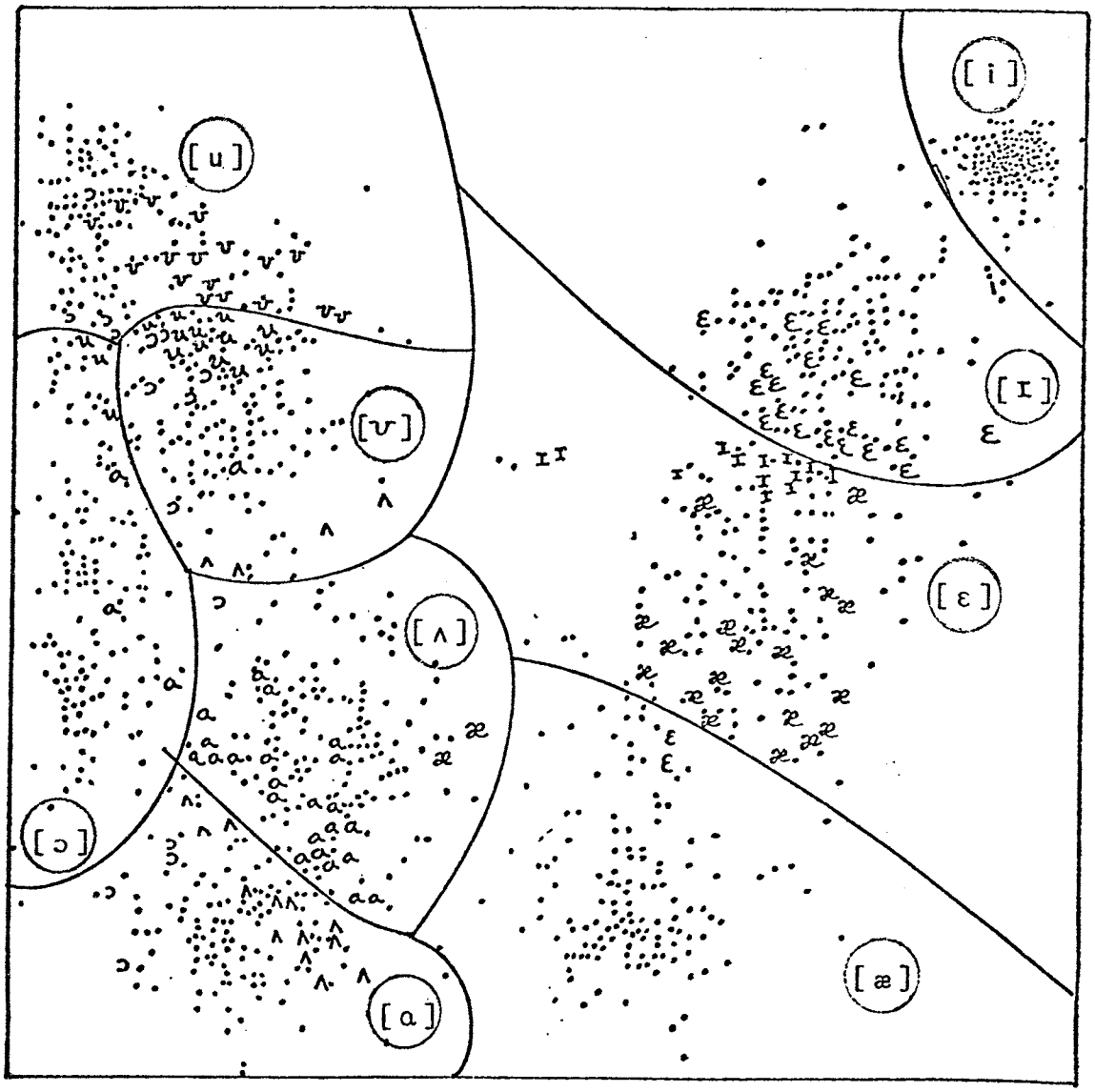


Figure 4.1.3. Gerstman normalization of P&B data.

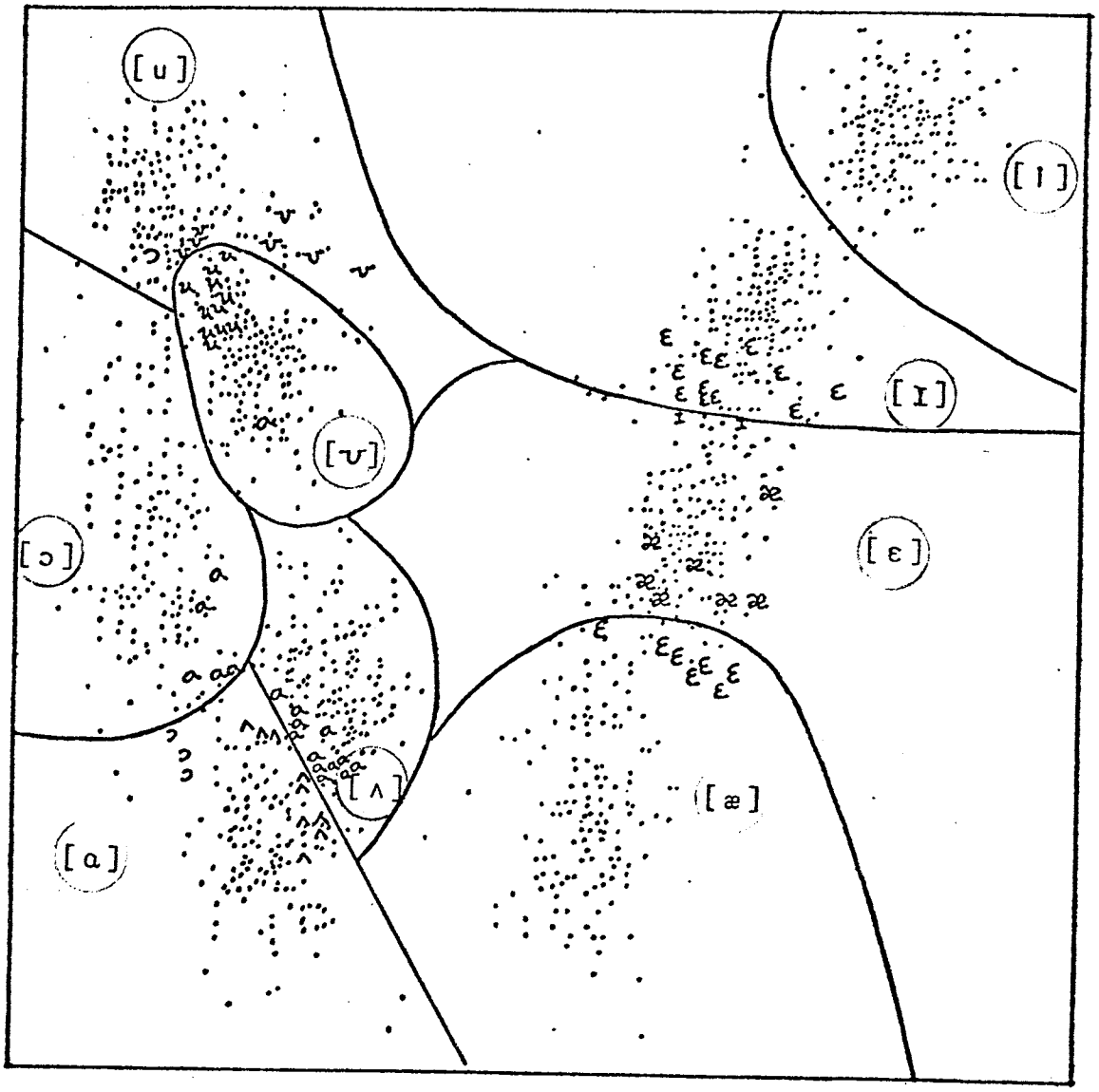


Figure 4.1.4. Lobanov normalization of P&B data.

VALUES ACCOUNTED FOR BY VOWEL GROUP MEMBERSHIP in a one-way analysis of variance of normalized formant values.

For the two formant case, we will use a coefficient of resolution based on Wilks' lambda (λ). According to Cooley and Lohnes (1971:227), lambda was introduced by Wilks as the multivariate generalization of eta-square. In the univariate case it reduces to

$$\lambda = 1 - \eta^2$$

The quantity $(1 - \lambda)$ will be used as the coefficient of resolution in the bivariate case.

Discriminant function classification.-- The final measure of resolution suggested here is a percentage correct identification on an objective classification rule. A versatile "canned" program for discriminant analysis in the SPSS package (Nie, Hull, Jenkins, Steinbrenner and Bent 1975) has been used for this purpose. The classification procedure provided by the SPSS program is included primarily as a check on the graphic analysis. The classification rule in the discriminant analysis program is to assign a point to the group for which its calculated probability of membership is greatest, assuming a multivariate normal distribution with equal covariance matrices for all groups.

Results of the analyses

Table 4.1.1 presents the results of the analyses described above. The most important point to be made is that any of these relative formant techniques provide excellent results in classification. The worst rate of identification is nearly 88%.[1]

Lobanov's procedure scores at least slightly better than any of the other techniques on all of the measures of resolving power provided. CLIH2 scores next best on the basis of both the graphic and SPSS classifications. The coefficients of resolution are somewhat mixed in this regard. The two-dimensional coefficient of resolution and the univariate coefficient for F2 show the two-dimensional Gerstman procedure to have coefficients of resolution second only to Lobanov's. However, for both classification techniques, and on the F1 coefficient of resolution, the Gerstman procedure ranks last after both CLIH and CLIH2.

Objections to Labonov's model.-- Though Labonov's procedure does seem to provide a greater degree of resolution for this data, there are nonetheless reasons to prefer the simpler, additive models. It should be remembered that Labonov's model requires

TABLE 4.1.1

RESULTS OF NORMALIZATION PROCEDURES

Procedure Name	Wilks' Lambda	2-D Coefficient of Resolution	% ID SPSS	%ID Graph	Coefficient of Resolution F1*	Coefficient of Resolution F2**
(1)	(2)	(3)	(4)	(5)	(6)	(7)
CLIH	.00704	.99296	87.72	90.28	.8900	.9370
CLIH2	.00542	.99458	89.91	92.90	.8995	.9453
Gerstman	.00471	.99529	85.31	89.25	.8867	.9609
Lobanov	.00269	.99731	91.23	94.30	.9090	.9724

more speaker-dependent information than the point normalization models. Though it is not entirely clear how to weight this in the evaluation of the procedures, there are at least two reasons to suspect that the Lobanov model of normalization is not sufficiently constrained if it is considered not simply as a recognition procedure but also as 1) a statement of the relationships that may obtain among the formants of the vowels of a single speaker and 2) a model of human vowel perception.

Concerning the first point raised above, it appears that the actual variation in formant frequencies of natural data is considerably more constrained than would be the case if vowels had ONLY to satisfy the conditions of Labonov normalization. The evidence for this is that the parameters extracted by the algorithm are not independent in the natural data. Rather, a fairly strong linear correlation exists between the two F1 parameters extracted for each speaker, the mean and standard deviation of that formant ($r = .7801$, $p < .01$). A similar correlation exists for the F2 parameters ($r=.8010$, $p<.01$). Interestingly, we would expect the means and standard deviations of formant values measured in hertz to be correlated if there was a log-interval relationship underlying the data. In fact, a linear correlation between means and standard deviations (or equivalently, between squared means and variances) is often taken to indicate that a log transformation is called for to produce additive models (cf. Winer 1971:400). Analyses of means and standard deviations in log transformed values indicate that there is no longer a significant correlation between these parameters in either G1 ($r=-.0228$) or G2 ($r=-.0017$).

The second reason for objecting to the Lobanov procedure is that additive point normalization models can account for the perceptual results of the experiments discussed in section 3.2 above, while range normalization procedures cannot.

CLIH and CLIH2

There is some reason to believe that only one subject-dependent parameter is free to vary in natural data and that CLIH rather than CLIH2 is a more appropriate model for normalization. The primary evidence for this from the analysis of the P&B data is that the two parameters of CLIH2, the average G1 and average G2 of each subject, are correlated ($r = .8499$, $p < .01$).

When this correlation is taken together with the weak evidence for the special status of [i] suggested in section 3.2, the case for the notion that speaker variation is really constrained along a single dimension seems to be strengthened.

Nonetheless, there are several reasons that have led us to use CLIH2 as the basic model for many of the analyses to follow.

First, CLIH2 does perform somewhat better in the normalization presented in this chapter. Second, graphic analyses presented in section 4.3, which are modifications of procedures used by Fant (1966, 1975), indicate that G1 and G2 behave somewhat differently in the range corresponding to about 700 to 1000 Hz. However, perhaps the strongest reasons for concentrating on two-parameter additive models (such as CLIH2) are the practical simplifications that result from the fact that F1- and F2-based data may be treated independently. While in the long run this may turn out to be an OVERsimplification, for the preliminary analyses to be presented here the advantages probably outweigh the disadvantages.

4.2. Two-way Analyses of Variance on Single Formant Data

The ANOVA model.-- For reasons outlined above, we will consider in this section the CLIH2 model which allows us to treat G1 and G2 data separately. The additive model of log formant relationships of CLIH2 implies that G1 measurements are analyzable into speaker-dependent and vowel-dependent components. For data that perfectly conforms to such a model, such as the synthetic "data" presented in the previous chapter (Table 3.1.1), we can "explain" all G1 values by the following linear model:

$$G_{1[V]s} = A_V + B_s + \mu$$

where μ is a constant, A_V is a vowel-dependent deviation from μ and B_s is a speaker dependent deviation from μ.

Such a perfect realization is not to be expected in the real world for two reasons: 1) There is a component of random variability in formant frequencies. Speakers will not always produce the same formant frequencies even on careful repetition of the same items. Furthermore, there will be random errors associated with the measurement process itself. 2) The log-additive model is probably not exactly correct for natural data. That is, there may exist systematic errors in the predictions of such a hypothesis above and beyond the random fluctuations.

These two additional sources of variation are included in the linear model for a complete two-way analysis of variance for G1 data:

$$G_{1[V]sr} = A_V + B_s + C_{Vs} + E_{Vsr} + \mu$$

The additional subscript r now appears to allow for replicated measurements, i.e. the inclusion of more than one token of each vowel for each subject. E_{Vsr} is the random component assumed to

be associated with each measurement. C_{Vs} is an "interaction" term. Its inclusion in the model is the formal representation of the error discussed in item two above.

The two-way analysis of variance presented below provides estimates for the amount of the total variation in G1 measurements which may be associated with each of the subscripted elements in the full linear model. The variation is thus attributed to the following sources: vowel dependent effects (A_V), subject-dependent effects (B_s), vowel-subject interaction effects (C_{Vs}) and "error" effects $E_{(Vsr)}$. The reader is referred to standard texts (e.g. Winer 1972) for details of factorial analyses of variance.

Analysis of G1

The ANOVA table.-- The results for the (fixed-effects) analysis of variance for G1 of the P&B data is presented in Table 4.2.1. The first five columns of this table constitute the standard analysis of variance table; the last two columns will be explained below. The magnitude of the F values (column 5) indicate that the vowel effects, subject effects and interaction effects are significant beyond the .01 level.

It is hardly surprising that the vowel effects are significant, since the standard test of significance involves the comparison of the estimated variance of vowel, subject and interaction effects to the estimate of the within cell variance (token-to-token variance or "error"). The claim of significant vowel effects is simply that differences in formant measurements due to differences in vowels are greater than those due to token to token variation.

Similarly, the fact that subject-dependent effects are significant amounts to the claim that log-formant frequency differences associated with differences in speakers are greater than those due to token to token variation.

If "significance" is the sole criterion to be considered, then the only really interesting result is that the interaction effects are also significant. This is a negative result from the point of view of the log-additive hypothesis since it indicates that there exist deviations from the predictions of a purely additive hypothesis that are significantly larger than those due to random variation.

TABLE 4.2.1

ANALYSIS OF VARIANCE TABLE (G1)

Source	Sum of Squares	D.F.	Mean Square	F	Prop. of SS_T	F'
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Vowels	156.521	8	19.565	4161.7	.7892	1434.13
Subjects	24.276	75	.32369	68.8516	.12241	23.726
Interact.	14.3014	600	.023857	5.07012	.07211	***
Within Cells	3.21563	684	.004701	***	.01621	***
Total	198.315					

Proportion of variation explained.-- However, it must be realized that the token to token "error" variation against which the various other effects are compared is extremely small. This is evident from the proportion of the total sum of squares SS_T^2, accounted for by the various effects in question. These proportions are presented in column 5 of Table 4.2.1. We see that within cell variation accounts for less than 2% of the total variation in the measurements.

Variation attributable to vowel differences amounts to about 79% of the total. About 12% of the total variation is attributable to speaker-dependent effects. While this amount may not appear very large, it represents about 58% of the variation remaining when vowel-dependent variation is removed from the total. Furthermore, the combined hypothesis effects (vowel effects plus subject effects) account for slightly more than 91% of the total variation.

The interaction effects constitute about 7% of the total variation. While this is about 59% as large as the amount of variation due to subject differences (and in that sense fairly substantial), there are several factors to be considered. First, it is reasonable to expect that some amount of this interaction is actually due to dialect differences among the subjects in question, since the additive hypothesis assumes phonetic equivalence in the vowels from different speakers. Secondly, the interaction effects have more "sources" than the vowel and subject effects. Column 6 in Table 4.2.1 provides F ratios that result from the comparisons of the variance estimates of the vowel and subject effects to that of the interaction effects.

Analysis of G2

Table 4.2.2 presents the results for a two-way analysis of variance of the G2 measurements of the P&B data. Once again, the main effects and interaction effects are significant beyond the .001 level.

Vowel effects account for about 85% of the total variation (Column 6). Speaker effects account for slightly less than 10% of the total variation, but this amounts to about 66% of the residual variation after vowel effects are taken into account. Vowel and speaker effects together account for slightly over 95% of the total variation. Interaction effects account for slightly more than 4% of the variation and token-to-token "error" for less than 1%.

The same general remarks that were made in connection with the G1 analysis also apply here. The fit of the additive hypothesis is quite good and interaction effects, though significant, are quite limited in scope.

TABLE 4.2.2

ANALYSIS OF VARIANCE TABLE (G2)

Source	Sum of Squares	D.F.	Mean Square	F	Prop. of SS_T	F'
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Vowels	199.560	8	24.9450	10840.7	.85394	2773.08
Subjects	22.5814	75	.301086	130.847	.09663	33.5572
Interact.	9.97617	600	.0166269	7.2259	.04269	***
Within Cells	1.57392	684	.0023015	***	.00673	***
Total	233.692					

Goodness of fit of the additive model

The overall results of the analyses just presented indicate that the degree of unexplained variation in the CLIH2 model is quite limited. The results of the categorizations presented in section 4.1 lead to the same conclusion. The CLIH2 model appears to provide an excellent first approximation to a mapping of acoustic parameters to phonetic features. In the next two sections, we will explore the problem of the nature of possible systematic departures from the log-additive hypothesis and we will suggest some steps that may eventually lead to the development of a still better additive model.

4.3 Criticism of Constant Ratios

Fant's criticism

The earliest relatively detailed criticism of the constant ratio hypothesis is provided by Fant (1966). In a graphic analysis of the AVERAGE formant data for male, female and child speakers of American English from the Peterson and Barney (1952) study and of similar data for male and female speakers of Swedish from his own study (Fant 1959), Fant argues that there are deviations from constant ratios in the comparison of the same vowels spoken by different speakers. Fant further notes that these deviations appear to depend on three factors: 1) the group membership of the speakers being compared (males, females or children), 2) the formant (F1, F2, F3) being compared and 3) the actual vowels in question. Fant argues that while the female-child comparisons (available in the American data only) appear to be relatively uniform, male-female comparisons exhibit patterns in the deviations from constant ratios that depend on the vowels being compared.

Fant suggests that these deviations may be related to differential pharynx to oral cavity length ratios in males and females. Though evidence for sex-linked physiological variation is limited, Fant argues that much of the deviation from constant ratios in formant data may be explained by systematic physiological differences superimposed on equivalent articulations in males and females.

Vocal tract analogue simulation of non-uniform vocal tract scaling

Most of Fant's physiological arguments were based on calculations of formant frequencies by assuming that certain articulatory configurations could be suitably approximated by ideal resonators. Recent vocal tract modelling of presumed male-female differences by Nordström and Lindblom (forthcoming) and Nordström (1975) do not corroborate Fant's (1966) hypothesis.

Nordström (1975) reports on experimental attempts to produce non-uniform ACOUSTIC scaling by means of the simulation of non-uniform ARTICULATORY scaling of different regions of the vocal tract. The simulated vocal tract modifications were applied to a set of "reference" area functions based on Fant's (1960) radiographic measurements of a male speaker of Russian. These reference area functions were divided "in the velar region" to allow for independent modifications of the "oral" and "pharyngeal" sections. Nordström reports the results of two separate series of experiments.

In the first series, only the relative lengths of the two regions were varied, since Fant (1966) had suggested that these were the most important aspects of male-female physiological differences. Though a wide range of scaling factors were examined, Nordström reports that there was remarkably little difference for the case of differential changes in the length of one of the regions from the case where the length was varied proportionately in both regions. That is, both uniform and non-uniform ARTICULATORY scaling resulted in essentially uniform ACOUSTIC scaling in F1 and F2, though slight non-uniformities were observed in F3 and F4. The acoustic results (for F1 and F2) were, then, essentially in line with the constant ratio hypothesis.

In a second series of experiments, cross-dimensional scaling was changed proportionately to length. While this did introduce some non-uniformity in the output, Nordström remarks that a plot of the formant ratios for corresponding vowels of different artificial speakers "... are not even remotely similar in F1 and F2"(1975:29) to corresponding plots for similar vowels presented by Fant (1966).

Implications for the failure of physiological explanations of scaling differences

Although this author is not aware of any explicit arguments along these lines, it seems likely that Fant's earlier explanations of deviations from constant ratios might have been interpreted as an example of "acoustic diversity from articulatory unity", an argument developed by proponents of the motor theory of speech perception particularly with regard to stop consonants. Two premises on which such an argument would have to be based are: 1) articulatory patterns in speakers of different vocal tract sizes are equivalent; and 2) non-uniformities in formant ratios across speakers are caused by biologically determined differences in the vocal tracts of different speakers.[3]

Even assuming the validity of point one above, Nordström's results indicate that point two does not account for deviations in natural formant data. Furthermore, if Nordström's experiments HAD corroborated Fant's hypotheses, there would still be serious reason to question the theory as a whole because there is no reason to assume that point one above holds. Indeed, the evidence presented in Chapter II of this work argues strongly against point one. Thus it would appear to be ill-advised to attempt to devise more elaborate accounts of the effects of biologically determined differences in vocal tract dimensions to account for non-uniform scalings between males and females, given the extent of variation that exists in individual patterns of vowel articulation.

But the failure to account for non-uniform relationships in acoustic scale factors in terms of an ARTICULATORILY invariant model does not detract in any way from the importance of such effects if they exist. If we are to understand the mapping between acoustic parameters and phonetic representations, such deviations may prove to be crucial.

Phonetic equivalence in the data sets

The detection of such non-uniformities and a delineation of their nature and extent is complicated by the question of the phonetic equivalence of the vowels of the different speakers (or groups of speakers) to be compared. Considerable dialect differences may exist within the two data samples considered by Fant (1966). Indeed, the Peterson and Barney data may contain a SYSTEMATIC bias related to the different groups as we are informed: "... Most of the women and children grew up in the Middle Atlantic speech area. The male speakers represented a much broader sampling of the United States; the majority of them spoke General American"(Peterson and Barney 1952:177). Nordström, after noting the difficulties with dialect in the American data, remarks that Fant's data for Swedish also includes speakers of differing geographical origins (Nordström 1975:21). The problem of dialect variation may affect all the data samples to be considered in this work.

Correlation of scale factors with formant frequencies

The hypothesis of systematic divergence in the ratios of male-female formant comparisons has been considerably strengthened in a recent study by Fant (1975). On the basis of graphic comparisons of average male and female formant values from six different languages, (and with reference to three other sources), Fant concludes that his earlier observations about the general patterns of male-female formant ratios are corroborated. The existence of consistent differences from such a variety of sources would argue strongly against the attribution of all such effects to accidents of regional dialect differences within samples, since there would be no reason to expect these to agree in different samples. While it is certainly possible that dialect differences might exist between males and females in any one language (as perhaps socio-linguistic markers) it would seem unlikely that any such differences would be consistent.[4]

In the absence of any definite evidence to the contrary, we will assume that the normal situation is for males, females, and children of the same speech community to produce phonetically equivalent sounds. With this in mind, any systematic differences from uniform scalings which hold over all languages may be assumed to preserve, and indeed may even be necessary to preserve, phonetic identity.

Fant (1975) presents a two-dimensional grid of scale factors for male-female changes which shows scale factors for F1 and F2 a number of points in the vowel space. He also provides evidence that improved normalization of average male-female values results when the scale factor grid is employed. While Fant's results are impressive, the normalization procedures he suggests are complex and the application of his scale factor grid appears to be somewhat imprecisely specified. In the present work, we will confine our investigations to modifications of the basic additive model presented first in Chapter III. The modifications we will explore are, however, closely related to one of Fant's observations about male-female scale factors, namely:

> There is a gross correlation between K_1 and F_1 and between K_2 and F_2. In other words, the percentage difference in F_1 and F_2 tends to increase with formant frequency (1975:4).

Seven grouped-data samples

The data for the analyses to follow overlaps considerably, though not entirely, with that considered by Fant (1975). It consists of average formant values from several distinct languages and dialects for male, female and (in some cases) child speakers. Table 4.3.1 indicates the source of the data, the number of speakers from each sex-age group and the number of vowels included in the analyses. The language and dialect is also specified, as well as a short code word which will be used on occasion to refer to the data sample in question. The first three of the samples listed in Table 4.3.1 (AMER1, AMER2, DANISH) include data from children as well as males and females.

While most of the following discussion will not depend on the comparison of vowel qualities across languages, a note on transcription is in order. With the exceptions noted below, the symbols employed are IPA transcriptions supplied either directly or indirectly[5] by the authors. The symbols "ü" and "ö" have been rendered by "y" and "ø" respectively; the vowel transcribed [a] in the UTRECHT data has been changed here to [ɒ] because it shows a lower F2 than the vowel transcribed as [ɑ]. The vowel transcribed as [a] by Frøkjær-Jensen has been denoted [æ] since the male formant values are considerably closer to those of American English [æ] than they are to either Swedish or Dutch [a].

TABLE 4.3.1

SEVEN SAMPLE GROUPED FORMANT DATA

Code	Language (dialect)	Male	Female	Child	Vowels	Source
AMER1	American English	33	28	15	9	Peterson and Barney 1952
AMER2	American English	5	5	5	9	Strange et al. 1976
DANISH	Danish	10	9	6	11	Frøkjær-Jensen 1967
DUTCH1	Dutch	5	5	*	12	van der Stelt et al. 1973
DUTCH2	Dutch	10	10	*	12	Koopmans-van Beinum 1973
UTRECHT	Utrecht (Dutch)	10	10	*	12	
SWEDISH	Swedish	7	7	*	14	Fant 1959

Displacement factor deviation plots

As noted in Chapter III, the constant ratio hypothesis is transformable into the constant log interval hypothesis. We find it convenient to continue to treat this model in its additive log form. Accordingly, the graphic techniques to be introduced here are aimed at portraying systematic deviations from the simple additive model. Under ideal conditions, we would expect all formant values of, say, the average female data measured in log-hertz to be equal to those of the corresponding male values except for a constant factor.[6] With real values, we might expect the values of the differences for the corresponding formants of corresponding vowels to fluctuate randomly about the average difference. This assumes that the average value of the differences between the log formant measurements is a good estimate of the hypothetical constant differences for the male-female comparison.

However, Fant's comment concerning a correlation between scale factors and formant frequencies would lead us to expect smaller male-female log formant differences at lower formant values than at higher. A useful tool for the investigation of such possible systematic bias from an additive hypothesis is what we will term a displacement factor deviation plot. Along the vertical axis, coordinate values represent deviations (differences) of the individual formant comparisons from the AVERAGE difference for the groups being compared. Formally, for a male-female comparison:

$$y = (G_{N[V]female} - G_{N[V]male}) - (G_{\cdot[\cdot]female} - G_{\cdot[\cdot]male})$$

As the horizontal coordinate value, we use the average of the two formant values in question:

$$x = (G_{N[V]female} + G_{N[V]male}) / 2$$

Similarly, deviation plots for female-child and male-child comparisons may be prepared when child formant averages are available.

Correlations of deviations with formant values

Since the deviation plots described above for all pairwise comparisons of speaker groups in the same language are all vertically centered about the zero deviation line, it is possible to superimpose the plots from different pairwise comparisons. Figures 4.3.1 and 4.3.2 are composite deviation plots for G1 and G2 respectively for all pair-wise comparisons of the seven grouped data samples. This plot includes male-female comparisons for all seven sets and female-child and male-child comparisons for AMER1, AMER2 and DANISH.

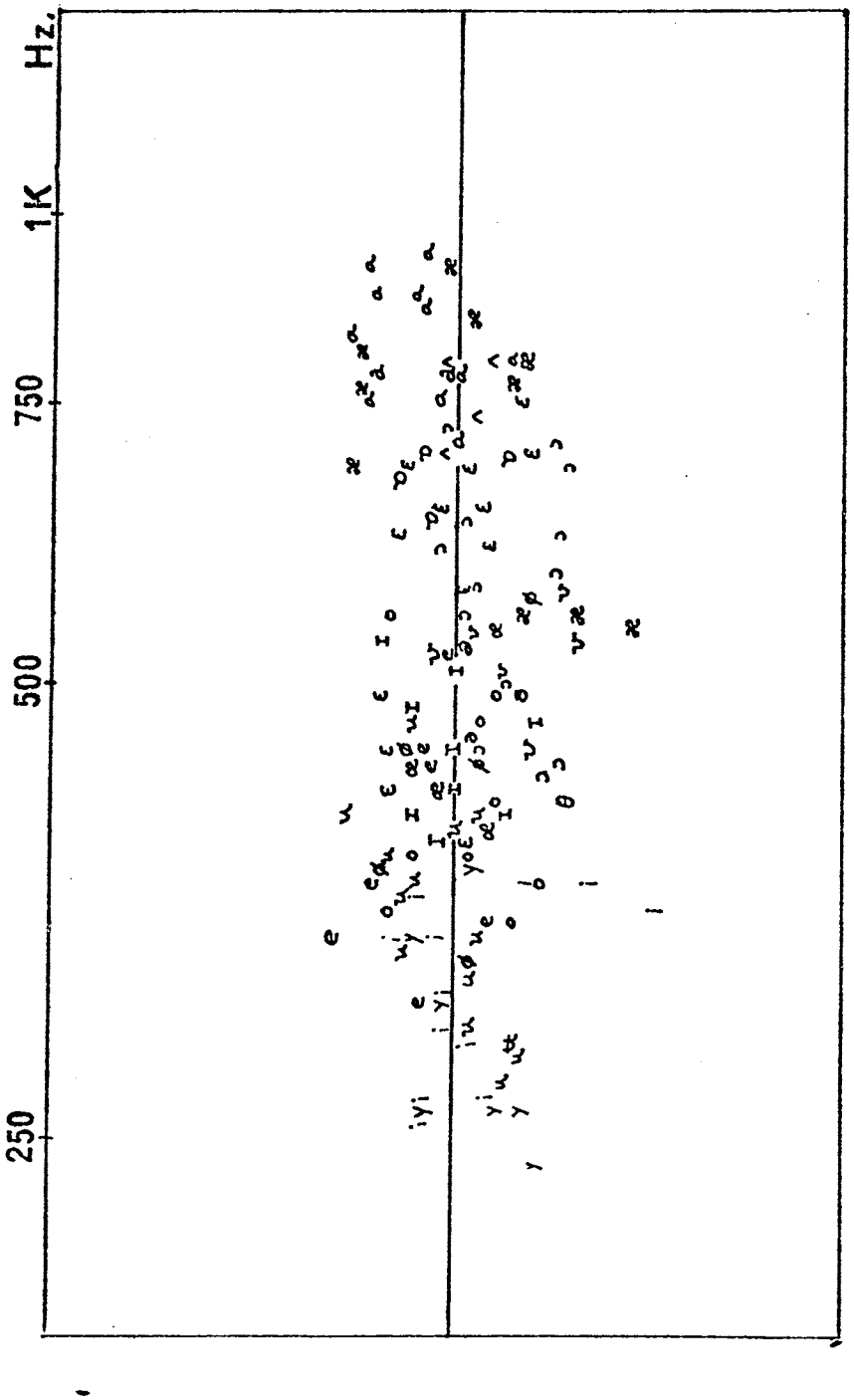


Figure 4.3.1. Displacement factor deviation plot for log(F1).

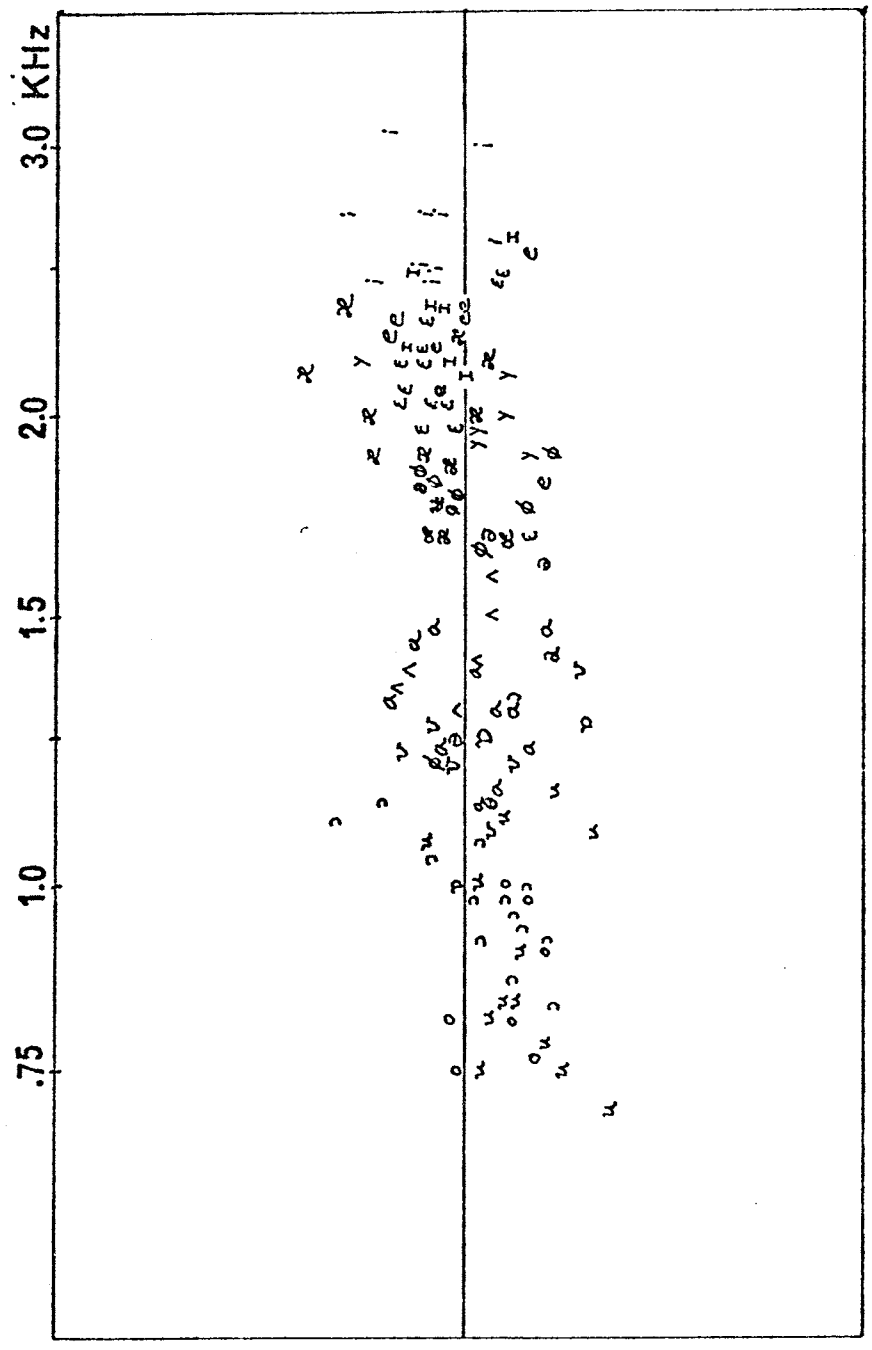


Figure 4.3.2 Displacement factor deviation plot for log(F2).

There are several tentative conclusions that may be drawn from an inspection of these plots. The horizontal center line represents the zero deviation from average difference; in effect, the expected value of all deviations under CLIH. Points below the line represent smaller than expected differences and points above the line larger. Figure 4.3.2 tends to corroborate Fant's claim of a correlation between scale factors and formant values in G2. Most of the points at low G2 levels fall below the average difference line and most at high values above.

There is little evidence of a frequency-related pattern in the G1 deviation plot (Figure 4.3.1). However, there does appear to be a relatively high proportion of greater than average differences in the range above 750 hertz.

By contrast, in the range between 750 and 1000 Hz, G2 deviations tend to be lower than average and in the G2 plot (Figure 4.3.2), the values are generally below the average difference line. Recall that the average difference for the pairwise comparisons in these plots is defined as the average over BOTH G1 and G2. It would appear that high G1 values and low G2 values are behaving somewhat differently even though the same general frequency range is involved.

Male-female, female-child, and male-child comparisons

As noted earlier, Fant (1966) has suggested that female-child comparisons may involve more nearly uniform scale factors than male-female comparisons. Fant's suggestion was based only on the average data from Peterson and Barney (1952). With a somewhat expanded data set, we find little suggestion that female-child comparisons differ in quality from male-female.[7]

Figures 4.3.3 to 4.3.5 represent respectively the displacement factor deviation plots for male-female, female-child and male-child comparisons for G1. Data from AMER1, AMER2, and DANISH are included. In this case, we are only considering G1 data in the calculation of the y-coordinate. That is, for a male-female comparison,

$$y = G_{N[V]female} - G_{N[V]male} - (G_{N[\cdot]female} - G_{N[\cdot]male})$$

There does not appear to be a simple linear trend in any of the three comparisons for G1. However, for all three comparisons there does appear to be a slight predominance of higher than average differences at higher F1 values. Female-child comparisons (Figure 4.3.4) do not appear to be appreciably more uniformly distributed about the average difference line than do male-female comparisons (Figure 4.3.3). The male-child comparisons (Figure 4.3.5)

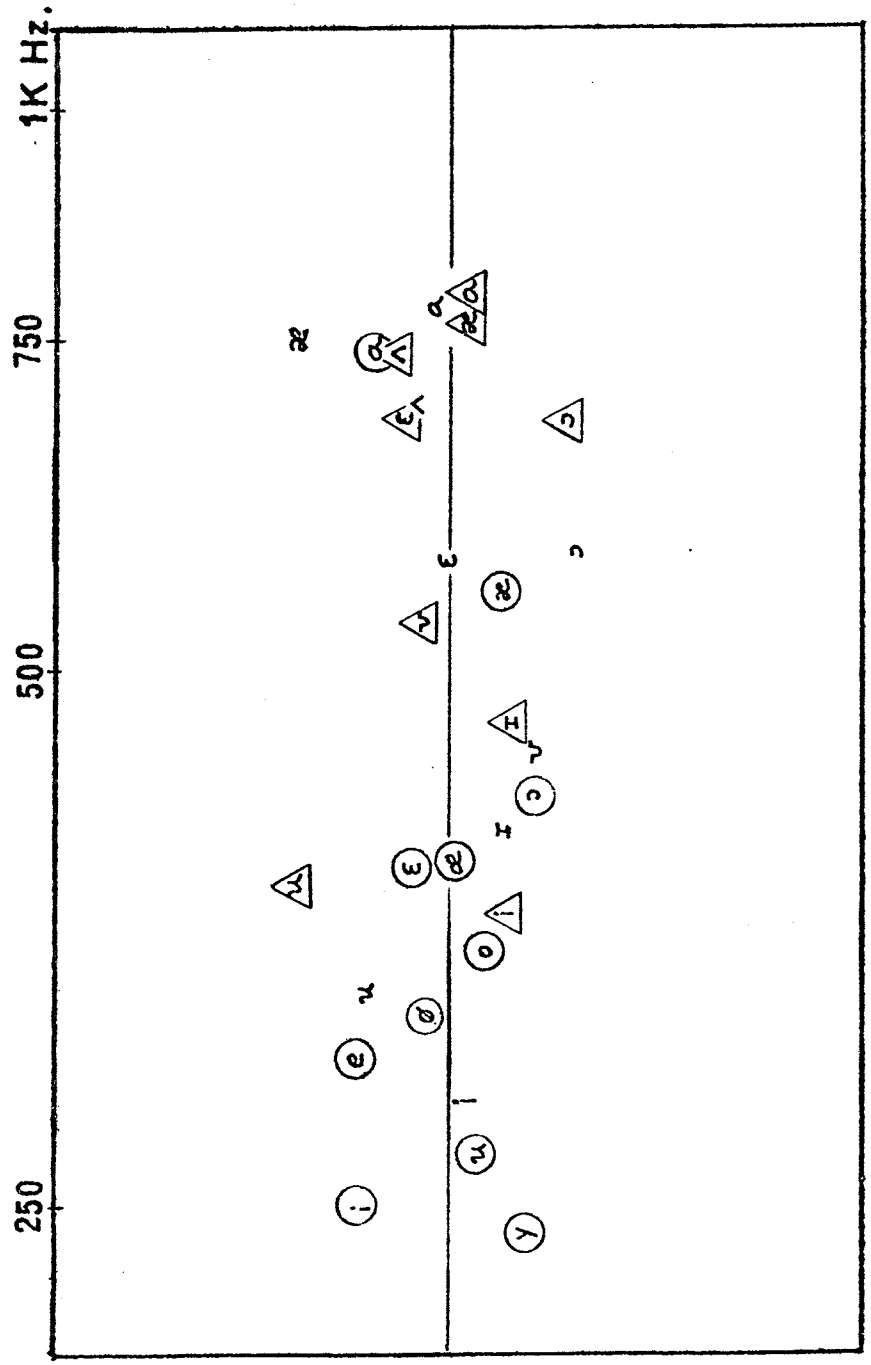


Figure 4.3.3 Male-female comparison for G1. Unmarked: AMER; △: AMER2; ◯: DANISH.

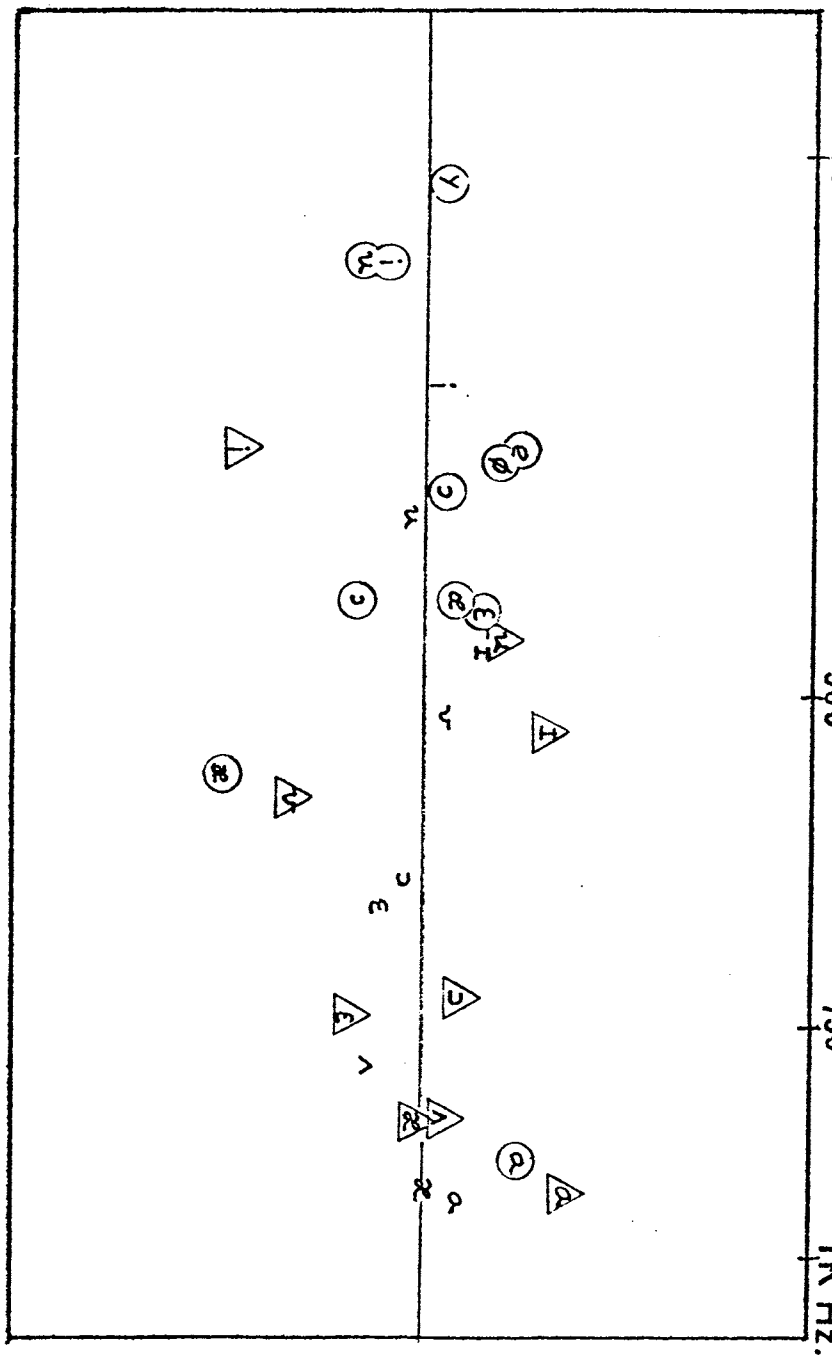


Figure 4.3.4. Female-child comparisons for Gl.
Unmarked: AMER1; △:AMER2; ◯:DANISH

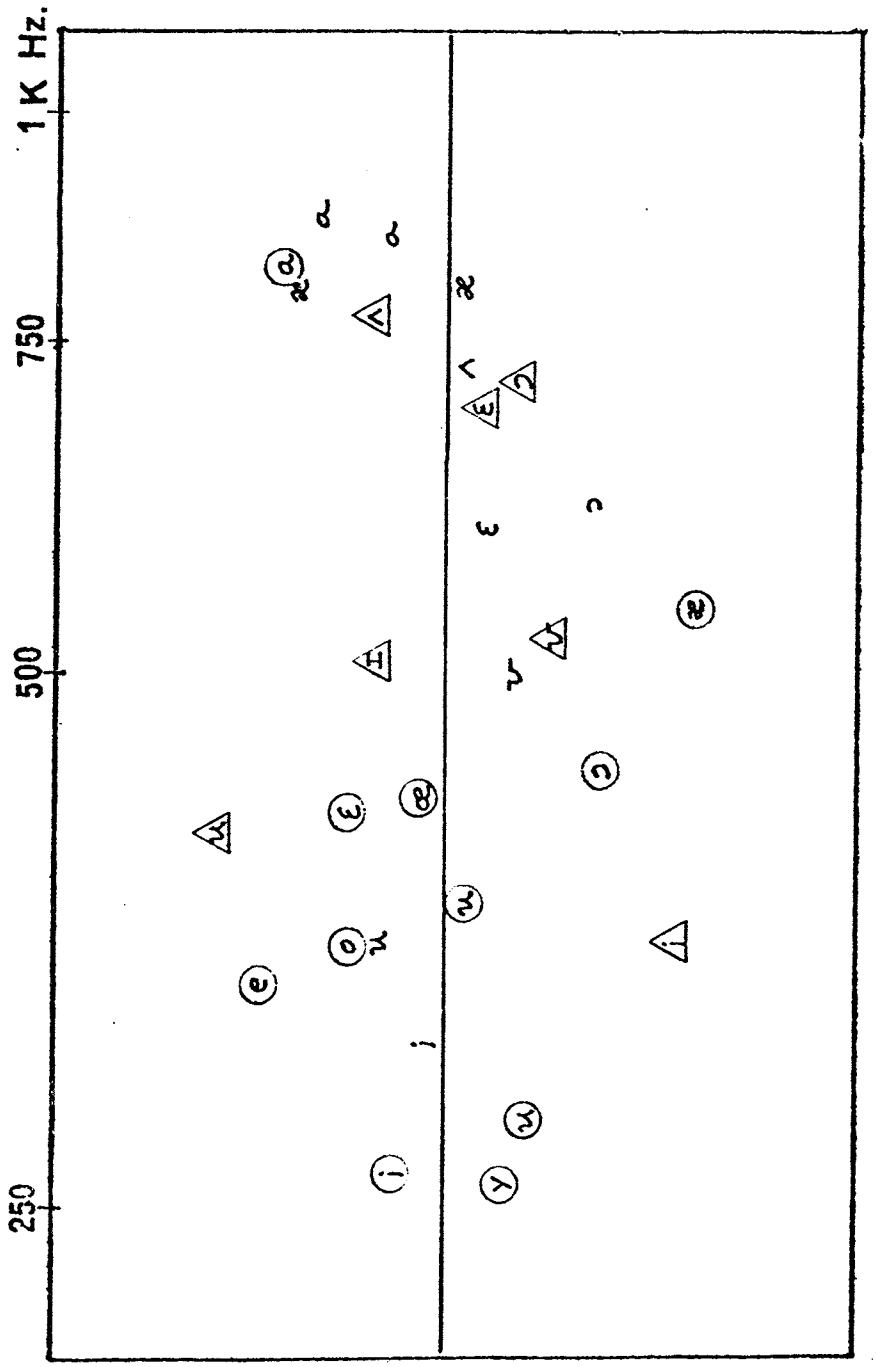


Figure 4.3.5. Male-child comparisons for G1.
Unmarked: AMER1; △ :AMER2; ◯ :DANISH.

show somewhat larger departures from the line, but there are no obvious trends in the directions of the deviations that are different from the other two comparisons. In Figure 4.3.6, the data from the preceding three figures are plotted together. In this combined plot, there is a weak indication of a roughly sinusoidal pattern in the deviations.

Figures 4.3.7 to 4.3.9 represent male-female, female-child and male-child comparisons for G2. The three pairwise comparisons seem to show the same trend: a positive correlation of the deviation with the value of the comparison. Figure 4.3.10 presents the combined data of the male-female, female-child and male-child comparisons.

4.4 Removable Non-additivity and "Voc"-transformations

In the graphic analyses presented above, there was weak evidence for some patterns in the deviations from simple additive models that was to some extent independent of the speaker class comparisons and even of the actual vowels involved. There may be a component of systematic bias in the log-additive hypothesis that is related SOLELY to the frequency values of the comparisons in question. The question arises whether it might be possible to reduce such systematic error by means of a transformation of the raw hertz formant measurements to a scale somewhat different from log. The estimation of a transformation to improve the fit of an additive model has been the topic of considerable discussion in the recent literature of statistics. For the rest of the chapter we will deal with F1 and F2 separately.

The method of Box and Cox

Considering first the case of F1, let us assume that there exists some monotonic function of F1, which we will call a "voc" function, in which transformed measurements for the same vowels of any two subjects are additive except for a random error component. In this case, the voc-additive hypothesis may be framed as a two way analysis of variance problem for which there is an error term, but for which the interaction term is assumed to be zero. The linear model for such a state of affairs may be stated thus:

$$V_{N[V]s} = A_V + B_s + E_{sV} + \mu$$

where $V_{N[V]s}$ is the voc-transformed F1 measurement; for F1 measurements A_V is a vowel-dependent component; B_s is a subject-dependent component; E_{sV} is the error term and μ is a constant.

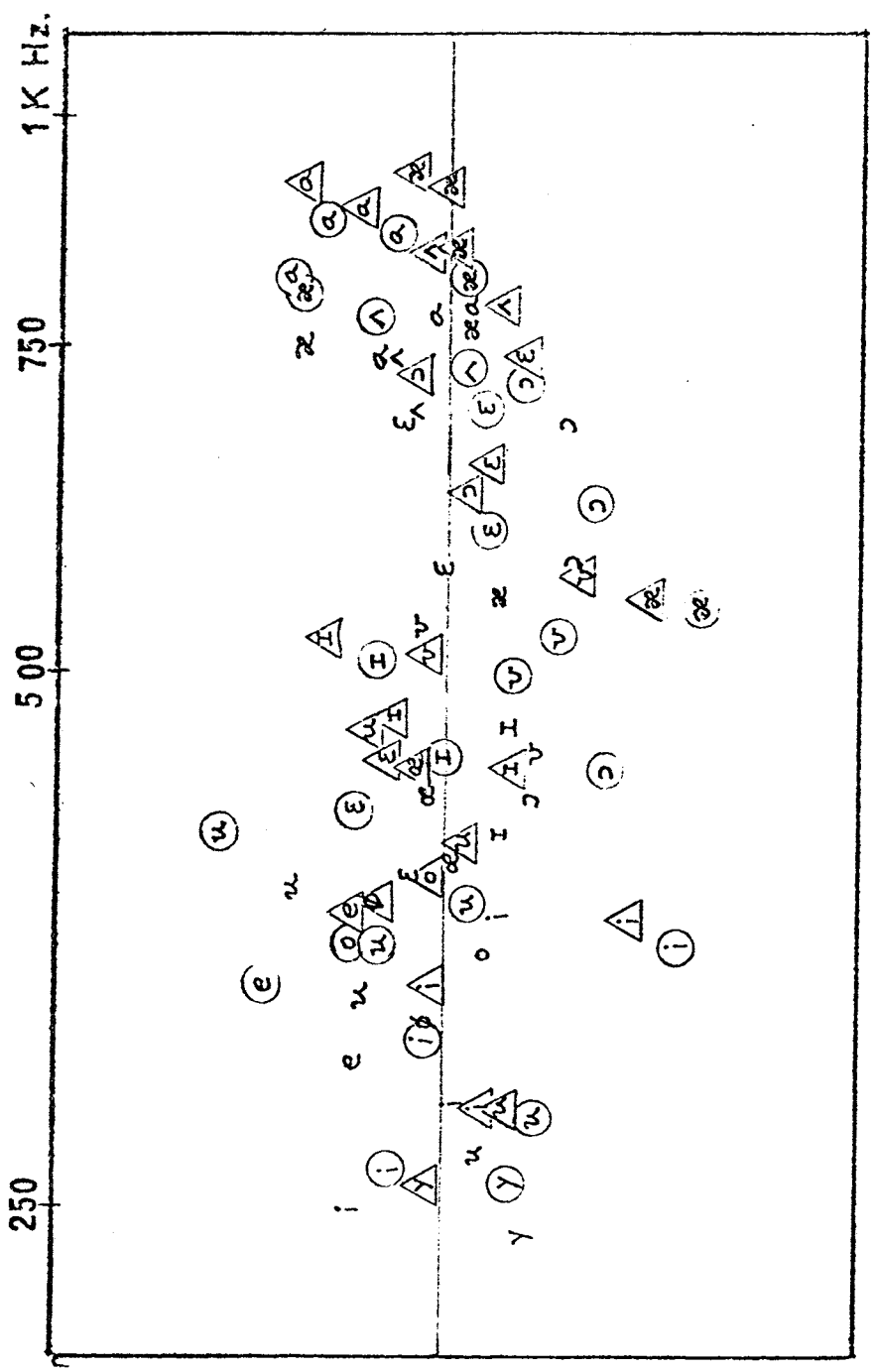


Figure 4.3.6. Unmarked: male-female; △: female-child; ○: male-child comparisons for G1. AMER1, AMER2 and DANISH data combined.

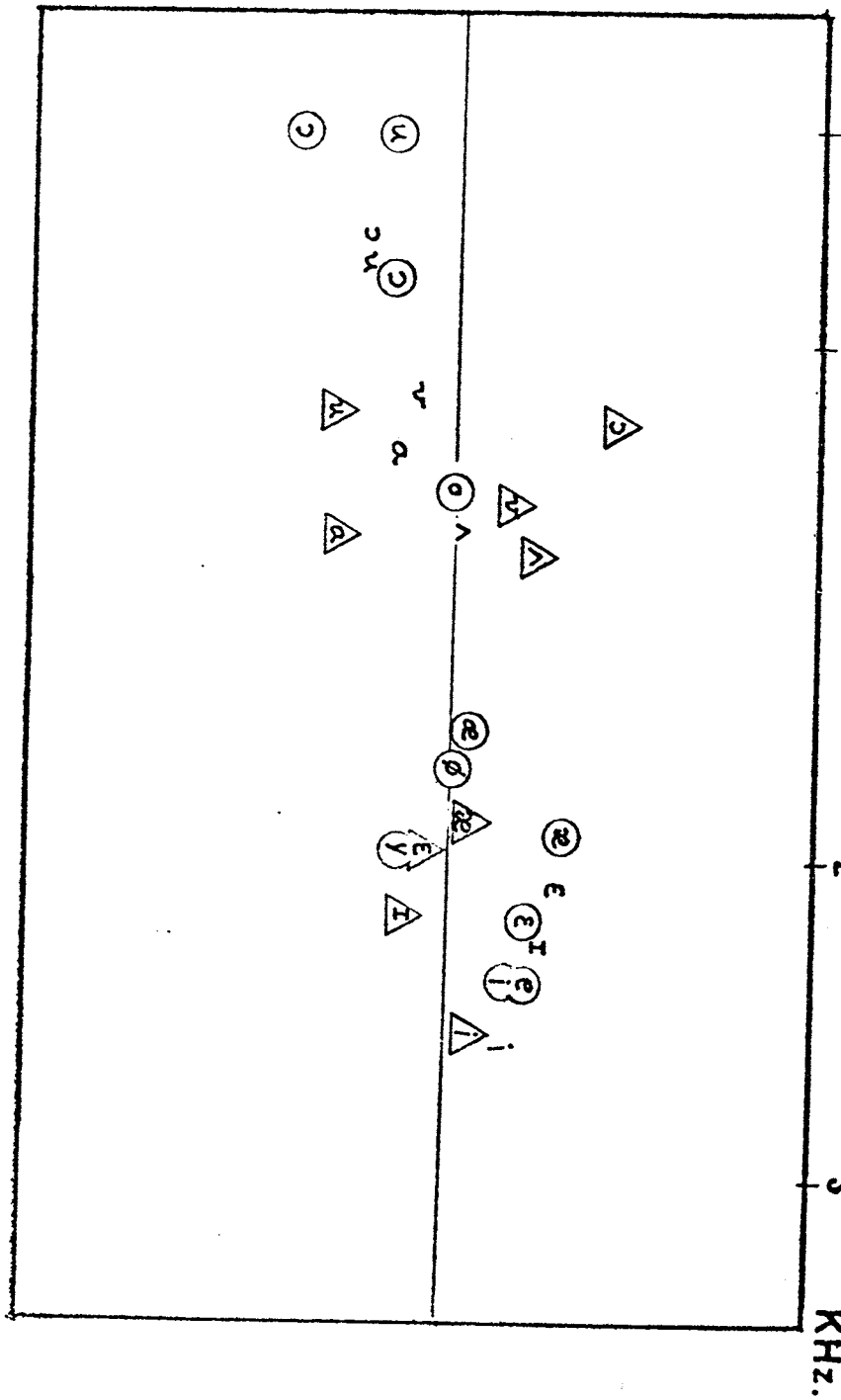


Figure 4.3.7. Male-female comparisons for G2.
Unmarked: AMER1; △:AMER2; ◯ DANISH.

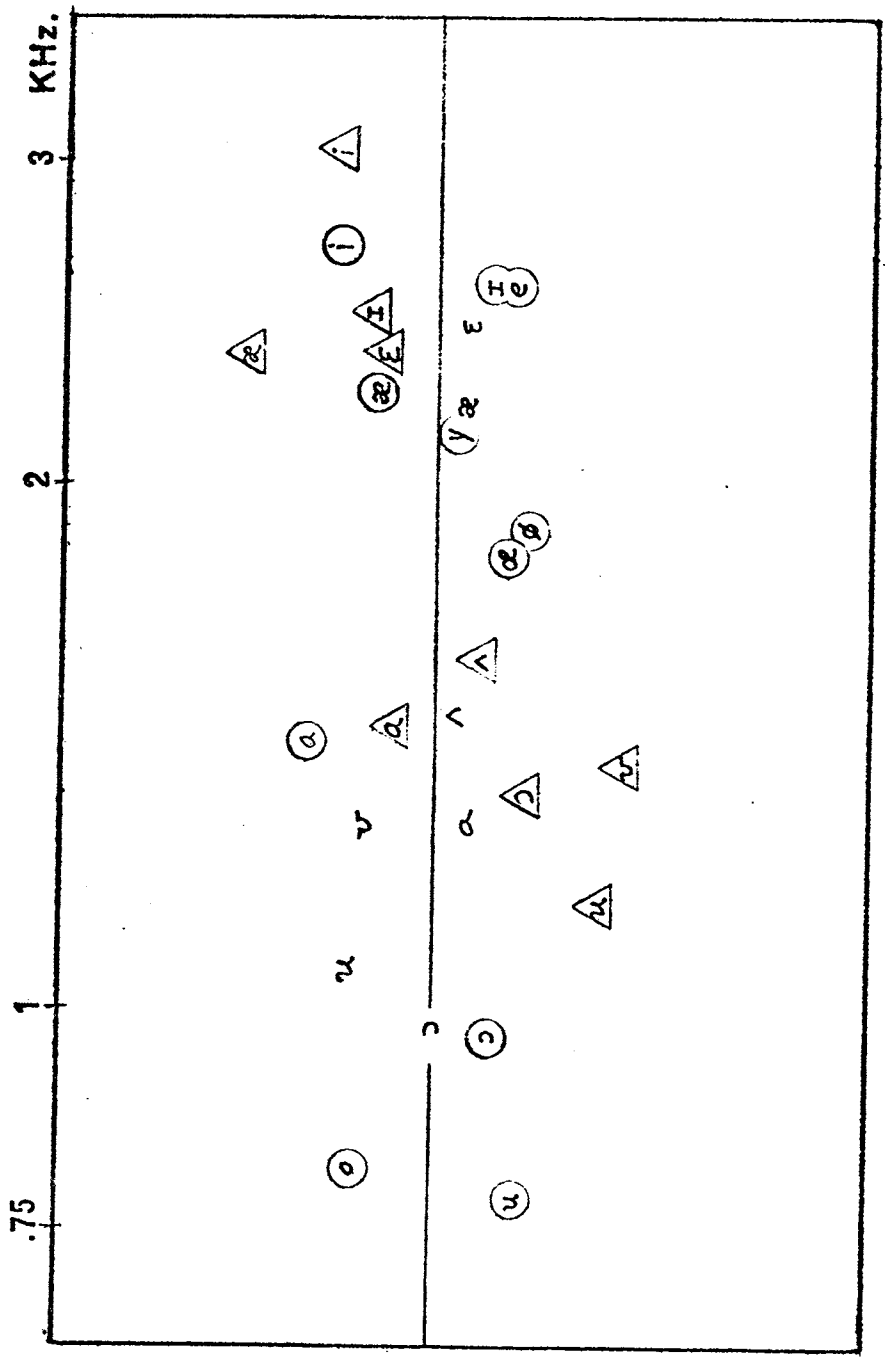


Figure 4.3.8. Female-child comparisons for G2.
Unmarked: AMER1; △ : AMER2; ◯ : DANISH.

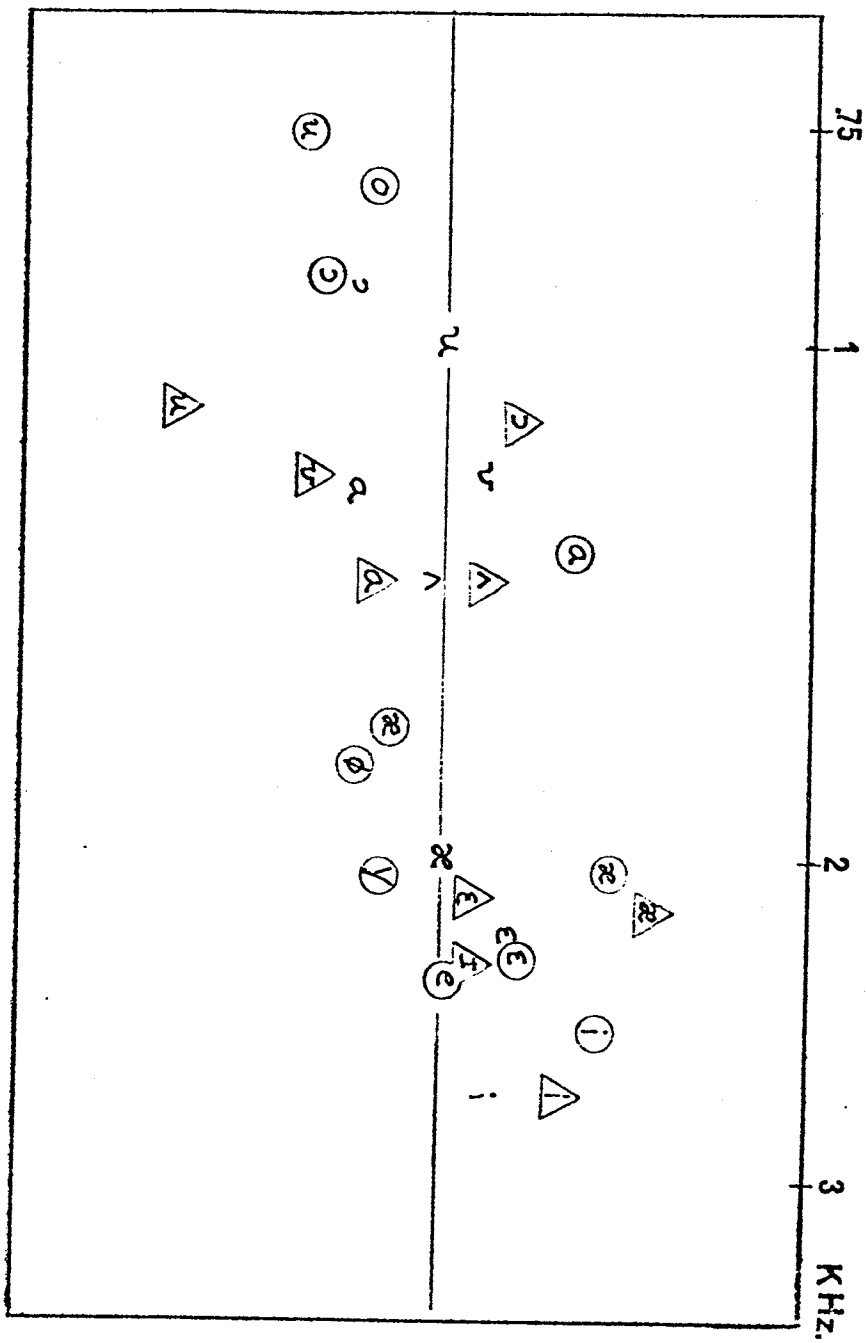


Figure 4.3.9. Male-child comparison for G2.
Unmarked: AMER1; △: AMER2; ◯: DANISH.

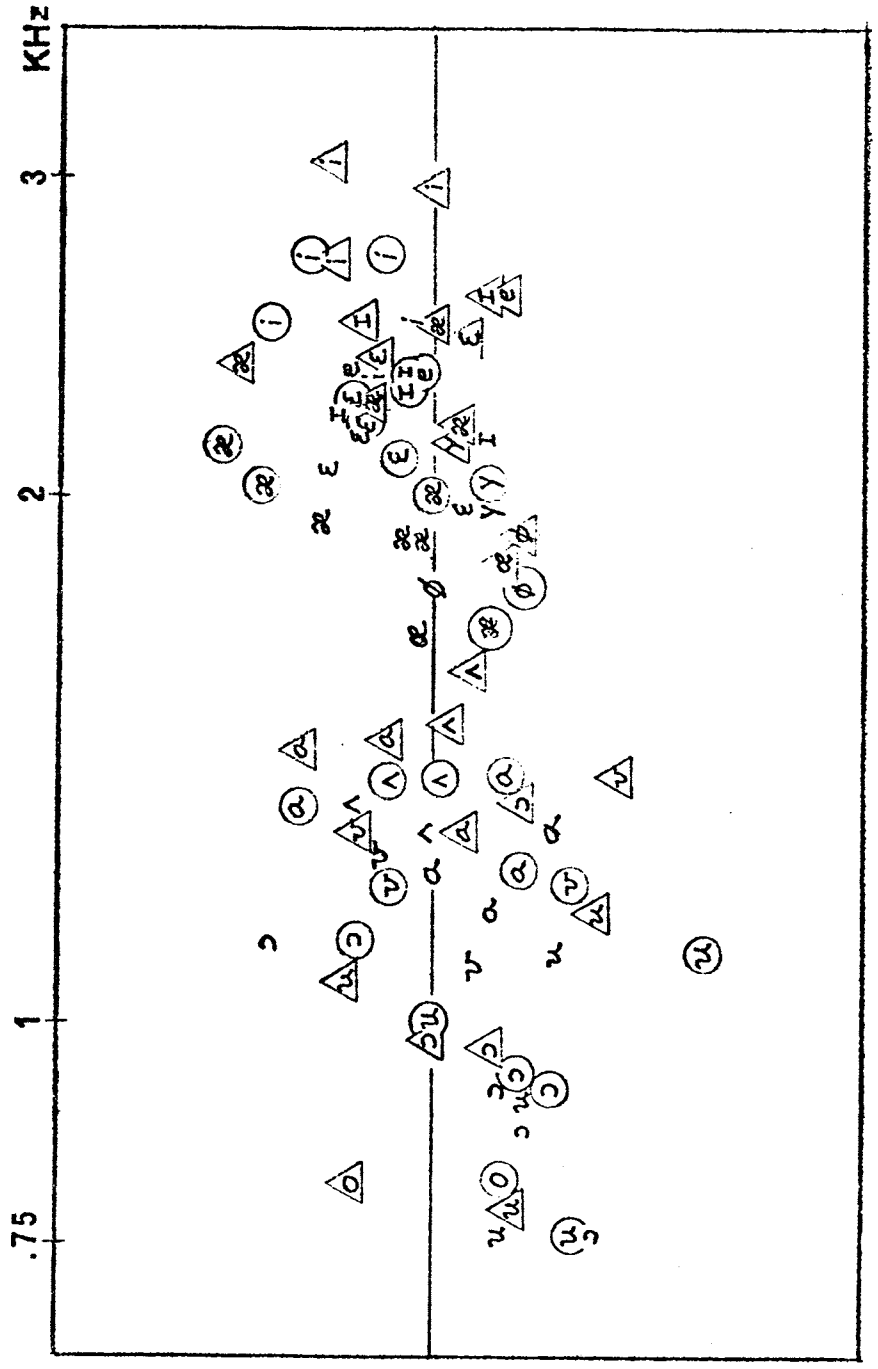


Figure 4.3.10. Unmarked: male-female; △ : female-child; ◯ :male-child comparisons for G2. AMER1, AMER2 and DANISH data combined.

If we assume that the error term is normally distributed with a mean of zero in the true voc scale, a method proposed by Box and Cox (1964) is applicable to our general problem.

Box and Cox propose a method for the maximum-likelihood estimation of the parameters (assuming normally distributed error term) in the family of transformations that may be specified by the following schema:

$$f(a,b,x) = \begin{cases} \log(x+b), & a=0 \\ (x+b)^a, & a\neq 0 \end{cases}$$

For the present analysis, it is convenient to base the transformations on formant values measured in kilohertz (kHz) rather than hertz. We will assume that the "true voc function" is a member of the family of functions specified by:

$$v(a,b,KF1) = \begin{cases} \log(KF1+b), & a=0 \\ (KF1+b)^a, & a\neq 0 \end{cases}$$

where KF1 is the F1 value measured in kilohertz. The technique suggested by Box and Cox for the two parameter case is to calculate the log-likelihood of the sample with respect to a range of parameter values (assuming a normal distribution of errors) and select those values that show the maximum log-likelihood. We can also look at other statistics of interest such as the coefficient of resolution in an analogous fashion.

Transformations of F1

In an investigation of F1, after some initial experimentation the log-likelihoods for all combinations of the parameters *a* and *b* were calculated, with *a* ranging from 1 to -2.2 in steps of -.2 and with *b* ranging from -.1 to +1.5 in steps of .1. The data from the seven grouped data sets were combined for this analysis (and most of those to follow) by treating the problem as a nested analysis of variance with both vowel and subject effects nested within languages. See Appendix II for the definitions of the sums of squares and the combined coefficient of resolution (the coefficient of resolution for the entire data set) involved.

The values in the range chosen resulting in the highest log-likelihood were a=-2.2, b=.9 (corresponding to the function $(KF1+.9)^{-2,2}$). Since this is on the border of the range investigated, the range was extended. However, several attempts at the extension of the range resulted in only very slight improvement of

the log-likelihood and no definite maximum was found.[8] The values a=-2, b=.8 showed nearly as great a likelihood as the values a=-2.2, b=.9. The former will be used in further discussion for convenience.

Table 4.4.1 (columns 2 to 4) presents a comparison of the function $(KF1+.8)^{-2}$ with linear and log analyses. This table shows the log-likelihood, relative log-likelihood (with respect to the highest three values) and the combined coefficients of resolution for these three functions. The linear function shows a much smaller log-likelihood than either the log function or the power function. The difference between the power function and log is not so great. We may convert log-likelihoods to likelihood ratios by taking the antilog of the differences between the two models to be compared. This gives a likelihood ratio of about 17.6 to one in favor of the power function.

Separate analyses of the seven grouped data samples indicated that the data set AMER2 showed relative peaks in the likelihood values at somewhat different values than the others. When the analysis was repeated on the other six sets combined, the peak of the likelihod surface for any given value of the parameter was found to be slightly shifted toward lower values of b. Table 4.3.1 provides the comparative statistics for the functions log (KF1) and $(KF1+.75)^{-2}$ columns 5 and 6). The difference in the log-likelihoods is somewhat larger for the six data sets than it was for the seven. It corresponds to a likelihood ratio of about 89 to 1.

Transformations of F2.

A similar procedure was carried out for F2 values of the seven grouped data sets. This time, a definite peak was found in the parameter range originally chosen for investigation. This peak occurred at values corresponding to the function $(KF2 -.2)^{-2}$ Since the value of the log-likelihood was only very slightly larger for this function than for the more convenient log (KF2-.4), the latter has been chosen for comparison linear and log functions. Table 4.1.1 (columns 7 to 9) presents the results of this comparison.

The relative gain for the modified log function, log(KF2-.4), in this case is considerable. The likelihood ratio is about 1.02×10^7. According to the procedures suggested by Box and Cox for establishing confidence intervals about the maximum observed log-likelihood, we would conclude that the log function is outside the .001 confidence interval for the parameters in question.

	F1 Seven Samples			F1 Six Samples		F2 Seven Samples		
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Function	KF1	log(KF1)	$(KF1+.8)^{-2}$	log(KF1)	$(KF1+.75)^{-2}$	KF2	log(KF2)	log(KF2-.4)
Log-likelihood	675.08	754.72	757.59	667.67	671.16	448.47	550.99	567.13
Relative log-likelihood	-82.51	-2.87	0	-4.49	0	-118.66	-16.14	0
Coefficient of Resolution	.9711	.9856	.9864	.9880	.9889	.9758	.9917	.9932

Table 4.4.1.
Log-likelihoods and combined coefficients
of resolution for grouped data samples

Graphic analyses of transformed values

It is instructive to examine the effects of the modified functions (the putative "voc" functions) on the supposed patterns in the displacement factor deviation plots. Figure 4.4.1 and 4.4.2 show the plots[9] for log(KF1) with that for $(KF1+.75)^{-2}$. Although there is little evidence for a pattern in the log plot, the power function plot seems to show a somewhat more evenly distributed (about the zero-deviation line) set of points. In particular, the predominance of greater than average deviations at high F1 levels that appears in the log plot (Figure 4.4.1) seems to be somewhat lessened in the power plot (Figure 4.4.2).

The modified F2 function provides a more noticeable improvement over log. Figure 4.4.3 shows log (KF2) and Figure 4.4.4 shows log (KF2-.4). The apparent correlation of deviations from constant differences with formant values is no longer evident in the deviation plot for the modified function.

Comparison with the P&B data set

If the modified functions discussed above represent true improvements in an additive model, we would expect that they would perform well on the individual P&B data used in the analyses presented in the first two sections of this chapter. Since modified functions for F2 appear to provide greater improvement in the grouped data analyses, we will discuss them first.

Transformations of F2 for P&B data. -- Because of the size of the data set, a limited search for an optimal funciton of the family log(KF2+b) was performed using the Box and Cox procedure, with b incremented in steps of .1 from -.5 to 1.3. The highest log-likelihood was found for b=+.8. This value is in fact "on the other side" of the simple log function (b=0) from the value selected from the analyses of the grouped data sets, where a value of b=-.4 was selected. The coefficient of resolution was found to peak in the P&B data at b=1.2. Interestingly (though probably accidentally) Fant's technical mel function is a linear transform of the function log(KF1+1) which lies between the functions with greatest likelihood and greatest coefficient of resolution for the P&B data.

The fact that the parameters selected for the grouped data and the individual data do not agree casts doubt on the apparent success of the attempt to find a "voc" function for F2. While this may be due to the inadequacy of a simple additive hypothesis, it is possible that an "effective second formant" (Carlson et al. 1970)

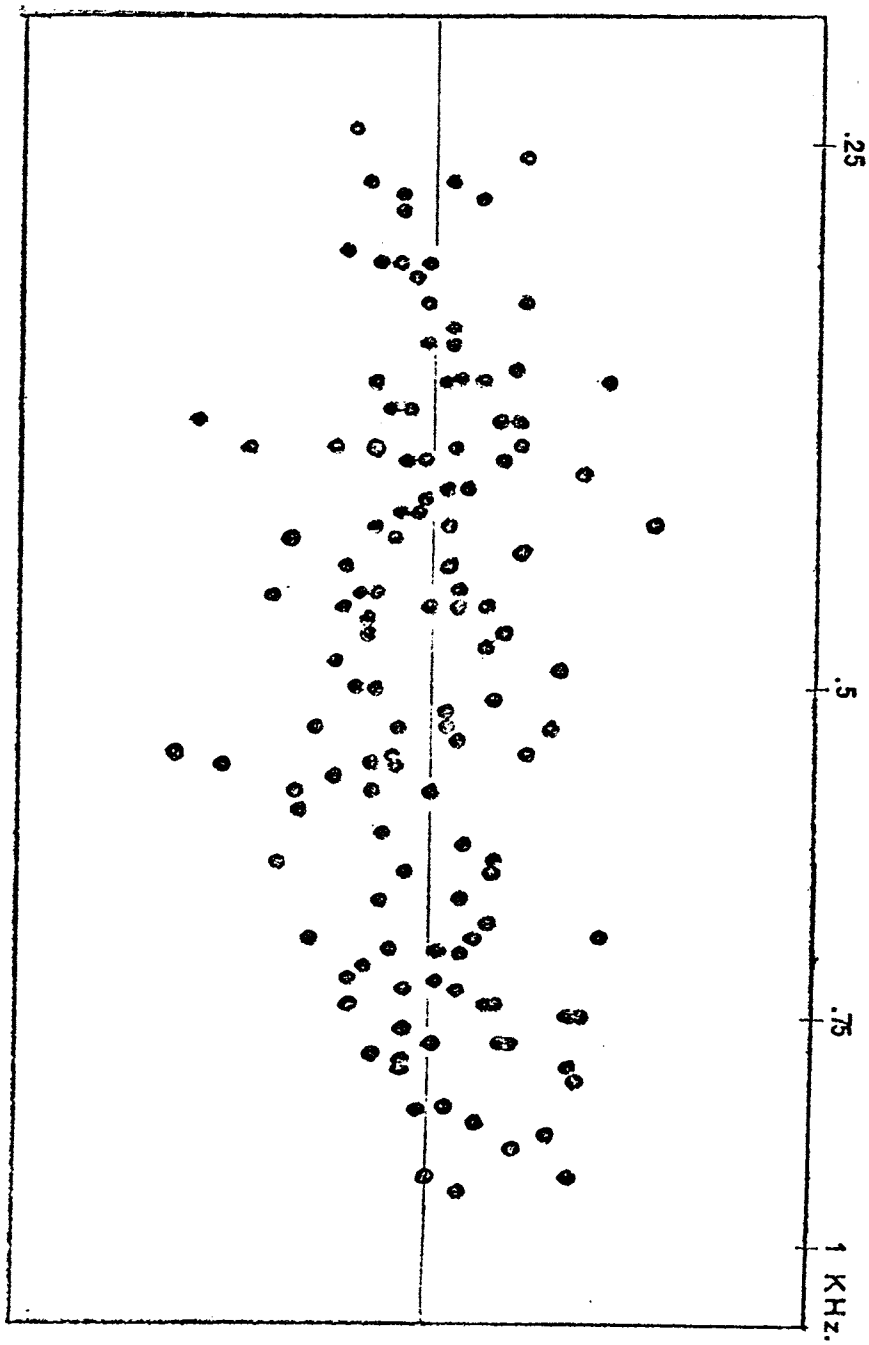


Figure 4.4.1. Displacement factor deviation plot for log(KF1).

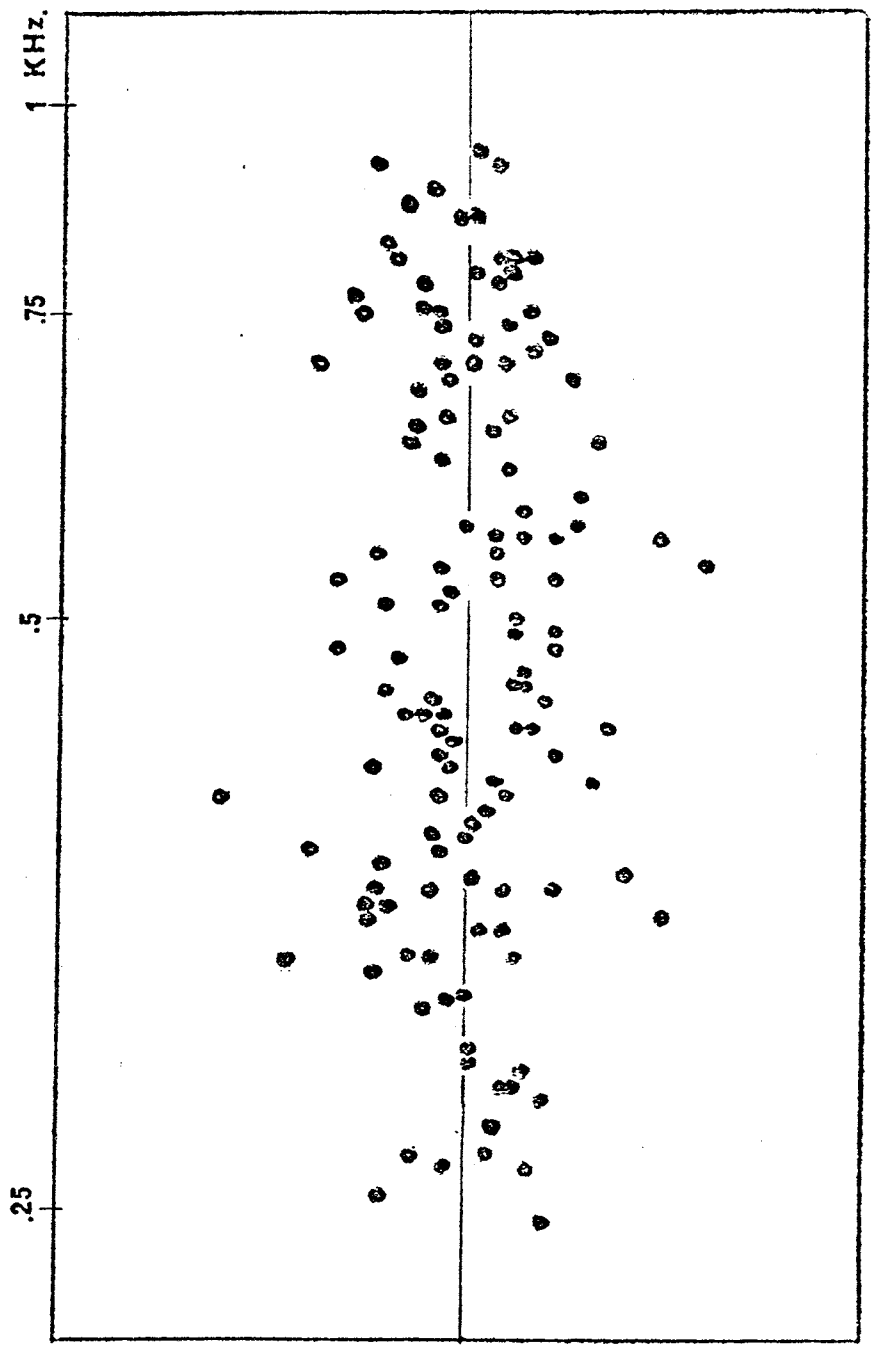


Figure 4.4.2. Displacement factor deviation plot for $(KF1+.75)^{-2}$.

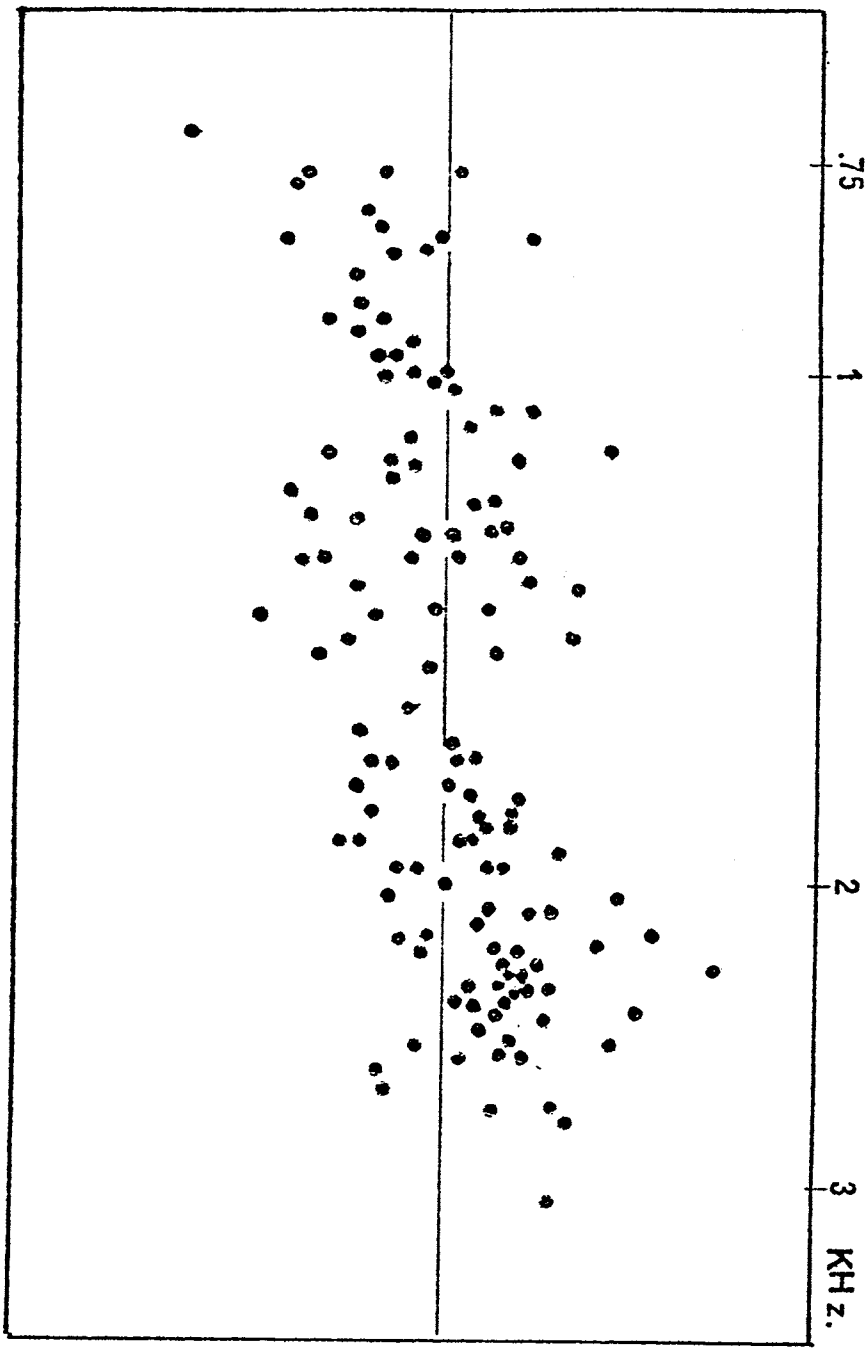


Figure 4.4.3. Displacement factor deviation plot for log(KF2).

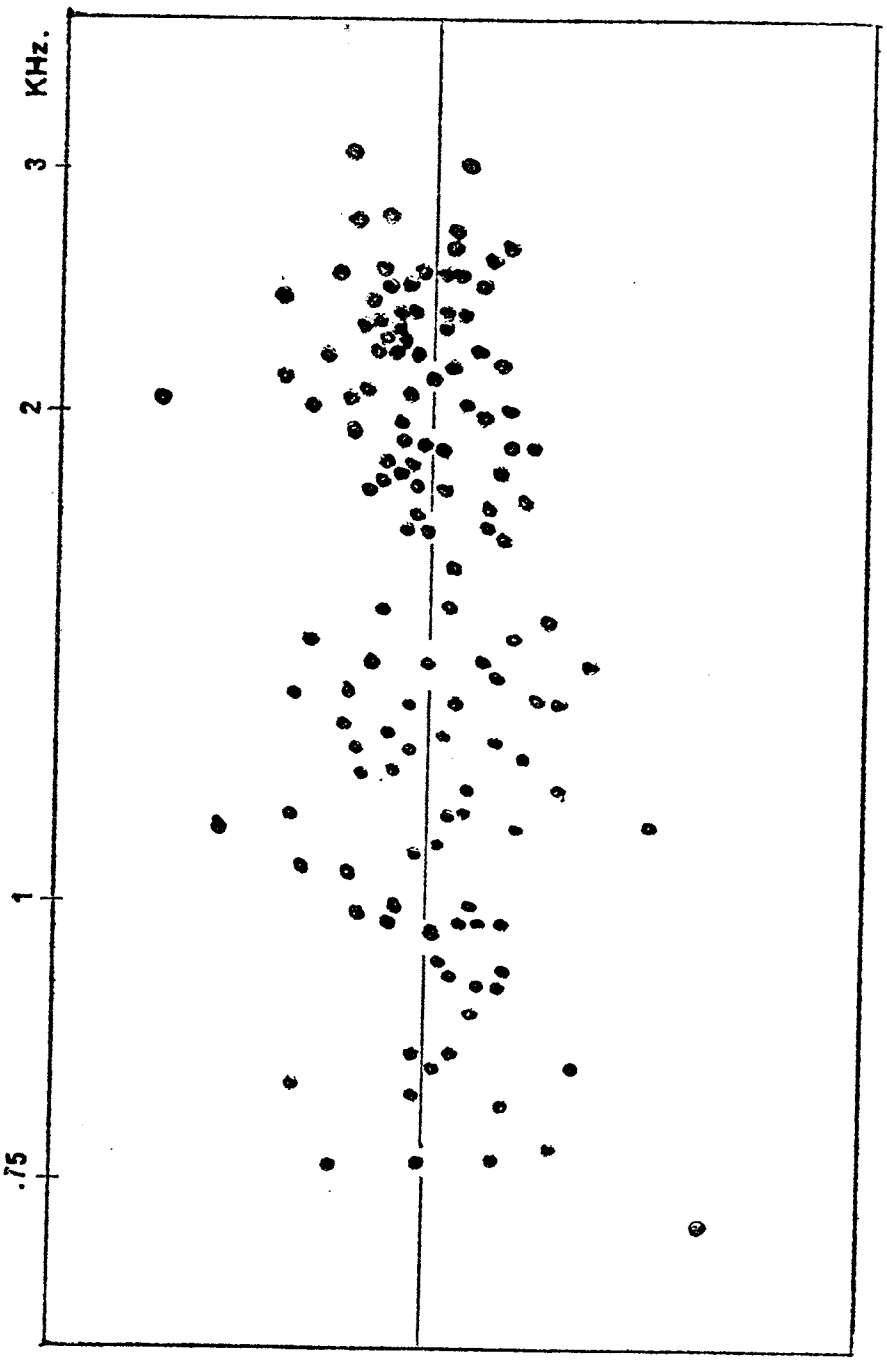


Figure 4.4.4. Displacement factor deviation plot for log(KF2-.4).

rather than F2 is the proper basis for the formulation of an additive model of the second perceptual dimension of vowels.

It is also possible that dialect variation and other phonetic variation such as diphthongization of some vowels, has resulted in the discrepancies between the grouped and individual data. The resolution of this issue must await analysis of "phonetically better-controlled" data.

Transformations of F1 for P&B data.-- In the case of F1, the analyses of the grouped data and the individual P&B data are somewhat more in accord. The family of functions specified by the formula $(KF1+b)^{-2}$ was investigated with the large data set with the P&B data with the parameter b incremented in steps of .1 from .4 to 1.8. The highest log-likelihood value was found at b=1.2. The likelihood as well as the coefficient of resolution is somewhat better for this function than for the log function. The maximum for the coefficient resolution occurred at b=1. The function corresponding to b=.8 (which was near optimal for the seven grouped-data sets) shows a slightly lower log-likelihood than the log function, but a slightly higher coefficient of resolution.

A modified power function for F1

Another approach to estimating a function for a better fit to an additive model has also resulted in a slightly better coefficient of resolution for the P&B data than any other function studied. Although the approach used is considerably less elegant and more ad hoc than the one described above, the function also behaves well on the grouped data sets.

The function is derived by a modification of a technique suggested by Anscombe and Tukey (1963) for the reduction of what they term "removable non-additivity" in a two-way analysis of variance. Their technique is aimed at a systematic deviation from an additive hypothesis that manifests itself as a correlation of the residuals of the additive model with the predicted values. When this correlation is "roughly quadratic" (1963:150), the provide a method for the estimation of an approximate power function for the reduction of the removable non-additivity. Tukey's one-degree of freedom for non-additivity is the appropriate test for the significance of the removable non-additivity.

For these analyses, the P&B data set was treated slightly differently than in the cases described so far. The data consisted of measurements of two tokens of nine vowels spoken by 76 subjects. Instead of using the 1368 measurements of the tokens, the two tokens for each vowel for each subject were averaged together and the analyses were carried out on the 684 individual averages.

When the Anscombe and Tukey analysis was performed on untransformed hertz values, there was a highly significant F for removable non-additivity and a power function estimate of -.21. However, when the analysis was repeated on log-transformed values, the F for removable non-additivity is not significant. An examination of a scatterplot of the residuals by the predicted values indicated the possibility of a somewhat more complex pattern.

The overall average of F1 values for the vowels is distributed in increasing order as follows: [i, u, I, U, ɛ, ɔ, ʌ, æ, ɑ]. The data was broken up into three overlapping ranges: low, [i, u, I, U], medium [I, U, ɛ, ɔ, ʌ], high [ɔ, ʌ, æ, ɑ]. Separate residuals analyses were performed on the three ranges with the log transformed values. While the middle range showed no significant non-additivity, both the high and low ranges did.[10]

This suggested an ad hoc procedure for the construction of a function with a higher degree of additivity by a piecing together of the three estimated power function modifications of the log function from the low, medium and high ranges.[11] The resulting function will be referred to as TEMPA.

An analysis of residuals was performed again with the measurements transformed by the TEMPA function. The general fit of the additive hypothesis (as measured by the eta-squared for the hypotheses effects) was improved for both the analysis of the entire set of vowels and also for each of the overlapping ranges. However, there was still significant non-additivity in the high and low ranges. The piecewise construction process was repeated using the TEMPA function as the base instead of log. This produced a second-pass pieced-together function, TEMPB. An analysis of residuals for the entire range and each of the overlapping sub-ranges for TEMPB transformed data now revealed no significant removable non-additivity.

A similar attempt at the piecewise construction of an F2 function was not pursued fully, since there was no indication of improvement after the first pass.

Rather than providing the actual formulae used in the calculation of the TEMPB function, tabulated values for a convenient linear transform of it are presented in Table 4.4.2.

Separate analyses of the seven grouped data samples

As a further test of some of the more interesting functions discussed in section 4.4, Tables 4.4.3 and 4.4.4 provide coefficients of resolution for each of the seven grouped data samples. Coefficients have been included for several additive normalizations as well as one-formant Gerstman and Lobanov normalizations.

TABLE 4.4.2

TABULATED VALUES FOR TEMPB FUNCTION

Hz.	TEMPB
100	53.5
150	127.4
200	200.0
250	276.4
300	352.8
350	428.7
400	513.4
450	586.9
500	658.6
550	728.5
600	796.5
650	853.1
700	904.9
750	946.6
800	980.6
850	1009.0
900	1033.0
950	1053.5
1000	1071.1
1100	1100.0
1200	1122.5
1300	1140.6
1400	1155.3
1500	1167.5

TABLE 4.4.3

COEFFICIENTS OF RESOLUTION FOR F1 FUNCTIONS
ON EACH OF THE SEVEN GROUPED-DATA SAMPLES

SAMPLE	KF1	log(KF1)	$(KF1+.75)^{-2}$	(TEMPB)	Gerstman	Lobanov
AMER1	.96568	.99040	.99134	.99303	.98543	.98728
AMER2	.96165	.97024	.96958	.97407	.96758	.97659
DANISH	.96226	.97931	.97989	.98100	.97313	.97791
DUTCH1	.98883	.99522	.99613	.99550	.99217	.99580
DUTCH2	.99110	.99536	.99463	.99652	.99576	.99585
UTRECHT	.98677	.98881	.98865	.98883	.98567	.98773
SWEDISH	.96553	.98634	.98868	.98471	.98632	.98028

	Hz	log	log(KF2-.4)	log(KF2+1)	Gerstman	Lobanov
AMER1	.97600	.99580	.99629	.99147	.99820	.99875
AMER2	.94619	.97508	.97571	.96901	.98980	.98295
DANISH	.97425	.99192	.99470	.98695	.98622	.99158
DUTCH1	.99534	.99667	.99700	.99623	.99620	.99342
DUTCH2	.98688	.99445	.99602	.99209	.99317	.99342
UTRECHT	.98575	.99431	.99558	.99194	.99421	.99404
SWEDISH	.97706	.99121	.99198	.98754	.99804	.99181

TABLE 4.4.4.

COEFFICIENTS OF RESOLUTION FOR F2 FUNCTIONS
ON EACH OF THE SEVEN GROUPED-DATA SAMPLES

Considering F2-based analyses first (Table 4.4.4), it may be seen that five of the samples show the highest coefficients of resolution on the function derived from the Box and Cox analyses, log(KF2-.4). In the other two cases, only the Lobanov procedure is better. From the point of view of the grouped data, the function log(KF2-.4) appears very attractive indeed. The failure of this function to produce better results on the individual P&B data is a disappointment.

Turning to the results for F1 summarized in Table 4.4.3, the function $(KF1+.75)^{-2}$ provides better coefficients of resolution than the log function on six of the seven samples. Note also that the TEMPB function derived from the individual P&B data provides a better coefficient of resolution than log function not only on the group averages from the same P&B data (AMER1) but also for five of the six independent data samples. TEMPB also provides better coefficients of resolution than Labonov's procedure on five of the seven samples.

Because of this relatively good performance of the TEMPB function on the independent data sets, a displacement factor deviation plot for this function is provided in Figure 4.4.5.

Graphic interpretation of selected F1-transformations

A convenient way to obtain some intuitive grasp of the general nature of the various transformations is to rescale them linearly so that a standard interval in hertz values corresponds to a standard (arbitrary) interval for each of the transformed values to be compared. This is possible since from the point of view of satisfying the requirements of an additive model, all linear transforms of any given function are equivalent to each other. The graphic analysis of three functions gives some indication that the TEMPB function and $(KF1+.75)^{-2}$ are to some degree "trying to do the same thing" in that there is some gross agreement in the deviation of their general shapes from that of the log function. See Figure 4.4.6. The two modified functions agree generally in that changes in hertz values at higher levels result in relatively smaller changes in the function values than does the log function.

Conclusions

While the results reported here cannot claim to have made substantial improvement over log-additive hypotheses, they do support some of the observations of Fant (1975) concerning possible bias in such a model. We have presented a number of analytic tools, both graphic and statistical that may prove useful in the

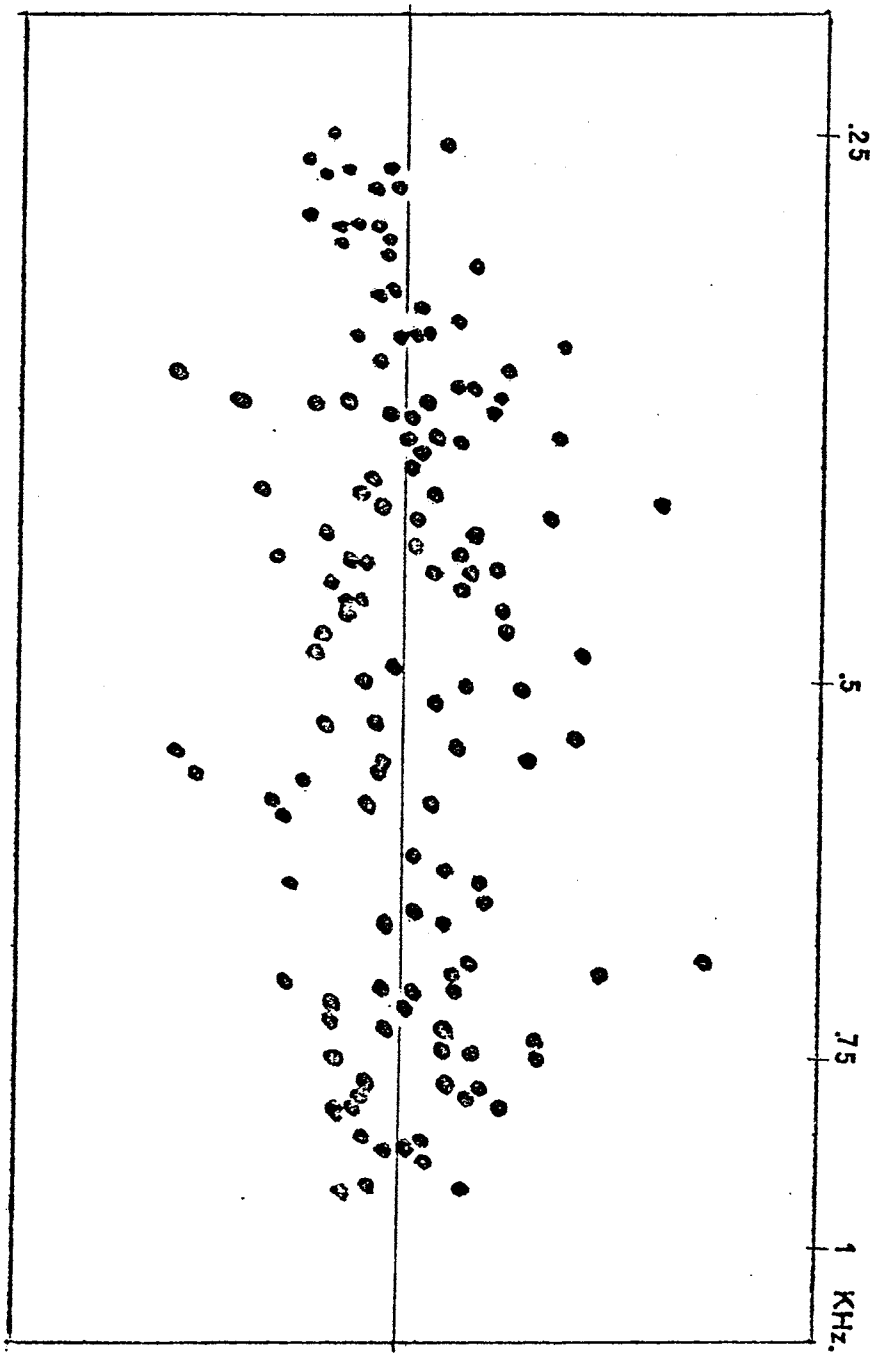


Figure 4.4.5. Displacement factor deviation plot for TEMPB transformed F1 values.

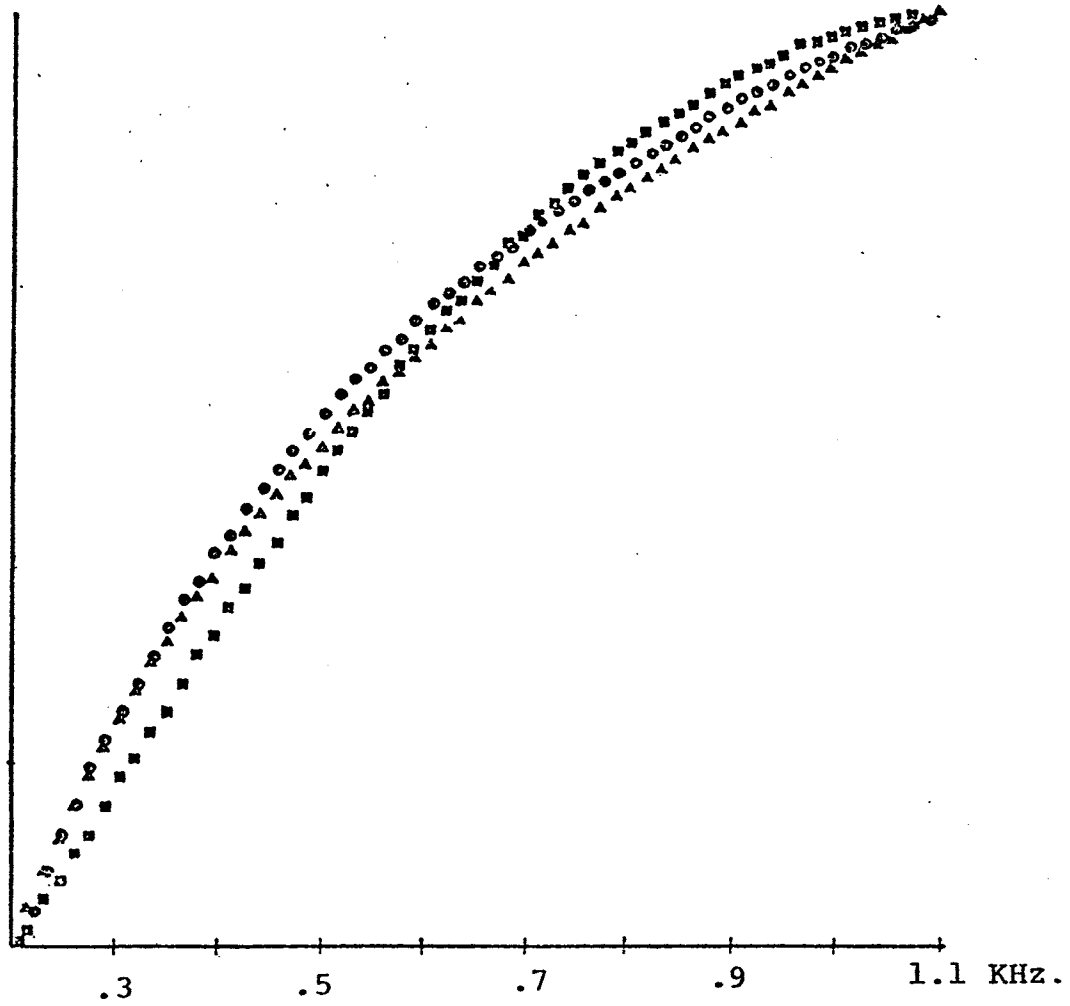


Figure 4.4.6. F1-function curves.

▲ log(KF1)

● $(KF1+.75)^{-2}$

■ TEMPB

ultimate formulation of a more adequate additive model of speaker differences. Since the fit of the CLIH2 model is quite good to begin with, we interpret the modest gains reported above as signs that such research should continue. It would appear that greater quantities of more carefully phonetically controlled data will be required if substantial progress is to be made. Generalizations of the analyses to include interactions of F1 and F2 and higher formants in connection with an "effective F2" would also seem worth pursuing.

Speaker-dependent variation in acoustic parameters appears to be quite systematic. On the basis of evidence presented in Chapter II of this work, this fact does not appear to be due to universal anatomical constraints imposed on an articulatorily invariant base; but on the contrary the systematicity of formant differences exists in spite of substantial idiosyncratic variability in the articulatory configurations assumed by individual speakers. Indeed, it seems reasonable to suggest that the primary link between the articulations of the same vowels by different speakers is that the acoustic consequences of such articulations be relatable to the output of other speakers by simple transformations not radically different from those investigated in this chapter.

Whether or not analyses of the type outlined above will prove correct, it appears that a simple point-normalized plot of formant frequencies results in a far more plausible mapping of phonetic features for vowel quality than any articulatory system thus far proposed. A theoretical breakthrough in articulatorily-oriented feature systems for vowels cannot be ruled out. However, at this point it seems likely that continued study of the problem of the sound-to-feature mapping in terms of unmediated functions of acoustic parameters holds more promise of success.

FOOTNOTES TO CHAPTER IV

[1]It should be noted that Gerstman's original normalization algorithm also included range normalized measures of the sums and differences of the first and second formants. The third formant was also rescaled, but it was used only in the separation of [ɜ]. Gerstman obtained an identification rate greater than any of the procedures described above (over 97%). However, it seems difficult to justify the potential extraction of ten speaker dependent parameters (maxima and minima of F1, F2, F3, the sum of F1 and F2 and the difference of F1 and F2) in light of the relative success of the point normalization procedures.

[2] This is what we have elsewhere referred to as the "total variation", defined as the sum of the squared deviations of each of the data points from the grand mean, $G_{1[\cdot]\ldots}$.

[3]Though it must be emphasized that such a position seems nowhere to be explicitly stated, the general philosophy of such an argument seems compatible with that of many accounts of "coarticulation effects", whereby essentially passive constraints on the vocal mechanism are held responsible for the diverse peripheral manifestation of an inherently invariant articulatory gesture.

[4]Fant remarks "...we cannot quite rule out the possibility of universal 'feministic' preferences in vowel qualities which might have influenced the average data." (1975:18) The present author shares Fant's apparent reluctance to do so, especially in the absence of any impressionistic comments to the contrary by practical phoneticians.

[5]Fant (1959) provides transcriptions in what he labels the STA alphabet. A conversion table for STA to IPA is provided in Fant (1967) with additional clarification in Fant (1969).

[6]Theoretically, we would expect some bias in averaged formant data because the averages are based on raw hertz measurements rather than on log-hertz. However an empirical comparison of geometric and arithmetic means for males, females and children computed on the (individual) P&B data indicate that errors introduced on such a basis are small enough to be safely ignored.

[7]Much of the appeal of Fant's arguments that female-child comparisons should be more uniform because of more uniform changes in vocal tract dimensions seems to be lost in view of Nordström's (1975) results discussed above.

[8]An extension of the range of the parameter *a* from about -5.0 brought about random results that suggested roundoff error. Several attempts were made to "delay" the roundoff error by increasing the precision of the program, and improving the efficiency of the calculations. None of these attempts provided a maximum value for the likelihood.

[9]These plots are based on single formant data alone. The deviation plot for modified functions are prepared in a manner analogous to that for the single formant log-based deviation plots discussed in the previous section. The horizontal scale is always marked with hertz values. The vertical scale is normalized in each chart so that the minimum and maximum deviations are a constant distance apart.

[10]This is the strongest indication for F1 data that there is a deviation from the log additive hypothesis that is related to the value of the formant frequencies in question. It is still something less than a "clear indication". However, an orthogonal polynomial regresssion of the residuals on the predicted values indicated a significant third degree correlation.

[11]This was done by selecting cutoff points at values of 400 and 600 Hz, which are near the midpoints of overlap between the low and mid, and the mid and high ranges, respectively. Linear constants were selected for each of the power functions so that a 50 Hz interval centered at the cutoff points mapped into the same values for the relevant adjacent "pieces" of the function. The resulting function was ascertained by graphic analysis to generate a reasonably smooth monotonic curve. See Figure 4.4.6.

APPENDIX I

PREDOMINANCE BOUNDARIES

The object of the predominance boundary plot used in Chapter III is to delineate smoothly bounded areas in the F1-F2 space for each vowel in which there is at least a 50% chance that a "primary conditions" response for that vowel will be given. This is done by estimating local boundary points for each vowel separately for F1 and F2. These local boundary points are connected by hand drawn curves. The method of local boundary point estimation is considered below.

Let us consider the case of establishing F1 boundaries for the vowel [x]. The F1-F2 stimulus-space is viewed as a row by column table. In each cell of this table, the responses are divided into two criterion classes: 1) those which are primary condition [x] responses, totalling X in number; and 2) all other responses, totaling Y. If X is greater than Y in a given cell, then [x] is said to predominate in that cell. To estimate F1 boundaries, the table is considered one F2 level (column) at a time. Starting in the column for the F2 level 1, the boundary algorithm first searches for the F1 level (row) in that column for which the value of X is highest. This is the mode of [x] in column 1.

If X is less than Y at the mode of the column, no boundaries are estimated in that column. Such a situation arises in cases where the algorithm is searching for F1 boundaries of a back vowel in high F2 columns: no F1 boundaries exist in that region of the F1-F2 space because the F2 level is inappropriate.

If there is [x] predominance at the mode, then a procedure searches first for upper then for lower boundaries of [x] predominance in that F2 column. In seeking the lower boundary point, the search procedure begins looking at each cell in turn, moving from the mode towards lower F1 levels in the column. (If the mode was found at F1 level 12, the search procedure examines levels 11, 10, 9, etc. in turn.) In each cell, [x] predominance is tested. If [x] no longer predominates (if Y>X), the present cell and the last cell examined are tested to see which is nearest the boundary; that is, for which of the two $|x/y - .5|$ is smaller. The "winner" of this test (which is presumably the one nearest the "true" boundary point) and the two adjacent cells to it are considered for the actual boundary point estimation process.

Using information in these three cells, two least-squares lines are estimated as approximations of the "skirts" of the presumed underlying categorization for [x] and non-[x] categories. Line A is the regression line of X values on F1 levels (in hertz) of the three cells and line B is the regression line of Y values on the same F1 levels. The intersection of the lines A and B is taken to be the lower predominance boundary for [x] on F1 in the F2 column in question.

A similar search is made outward from the mode towards higher F1 levels to determine the upper predominance boundary for [x] in that column. This process is repeated for all F2 columns.

An analogous process is used to determine F2 predominance boundaries for each F1 row. The boundary points output by this process appear to lead to satisfactory results for the pooled data. In most cases, the F1 and F2 boundary points "mesh" well and predominance areas are easily drawn by hand.

APPENDIX II

NESTED ANALYSIS OF VARIANCE FOR COMBINED DATA FROM SEVEN GROUPED SAMPLES

The linear model for this analysis may be stated as follows:

$$Y_{vs\ell} = A_v + B_s + C_\ell + E_{vs\ell} + \mu$$

$Y_{vs\ell}$ is the formant value of the v-th vowel of the s-th speaker class of the l-th language; A_v is the additive contribution of the v-th vowel of the l-th language; B_s is the additive contribution of the s-th speaker of the l-th language; C_ℓ is the additive contribution of the l-th language; $E_{vs\ell}$ is the error term; and μ is a constant.

Defining:

L = total number of languages

V = the number of vowels in language ℓ

S = the number of speaker-classes (males, females, children) in language l.

We may define the relevant sums of squares as follows:

Total sum of squares (SS_T)

$$\sum_{\ell=1}^{L} \sum_{v=1}^{V_\ell} \sum_{s=1}^{S_\ell} (Y_{vs\ell} - Y...)^2$$

Sum of squares for vowels within languages ($SS_{V(\ell)}$):

$$\sum_{\ell=1}^{L} \sum_{v=1}^{V_\ell} (Y_{v.\ell} - Y..\ell)^2$$

Sum of squares for speaker-classes within languages ($SS_{S(\ell)}$):

$$\sum_{\ell=1}^{L} \sum_{s=1}^{S_\ell} (Y_{.s\ell} - Y_{..\ell})^2$$

Sum of squares for languages (SS_L):

$$\sum_{\ell=1}^{L} (Y_{..\ell} - Y_{...})^2$$

The error sum of squares may then be defined as follows:

$$SS_E = SS_T - SS_{V(\ell)} - SS_{S(\ell)} - SS_L$$

The coefficient of resolution may be defined as:

$$1 - SS_E$$

This is equivalent to a weighted (proportionally to the number of data points in the sample) average of the individual coefficients of resolution of each language.

REFERENCES

Abramson, A. 1962. *The Vowels and Tones of Standard Thai: Acoustical Measurements and Experiments.* Indiana University Research Center in Anthropology, Folklore and Linguistics, publication 20. Bloomington: Indiana University Press.

_____. 1972. Tonal experiments in whispered Thai. *Papers in Linguistics and Phonetics to the Memory of Pierre Delattre.* The Hague: Mouton. pp. 31-44.

Albright, R. W. 1958. *The International Phonetic Alphabet: Its Background and Development.* Indiana University Research Center in Anthropology, Folklore and Linguistics, publication 7. Bloomington: Indiana University Press.

Anscombe, F., and J. Tukey. 1963. The examination and analysis of residuals. *Technometrics* 5.141-160.

Atal, B. 1974. Towards determining articulator positions from the speech signal. *Preprints of Speech Communication Seminars*, Vol. 1.1-9. Stockholm: KTH.

Bell, A. 1867. *Visible Speech.* London: Simpkin and Marshall.

_____. 1897. *The Science of Speech.* Washington: The Volta Bureau.

Box, G., and D. Cox. 1964. An analysis of transformations. *Journal of the Royal Statistical Society* B 26.211-252.

Broad, D., and R. Fertig. 1970. Formant frequency trajectories in selected CVC nuclei. *JASA* 47.1572-1582.

Broadbent, D., and P. Ladefoged. 1960. Vowel judgment and adaptation level. *Procedings of the Royal Society* B. 151. 384-399.

Brücke, E. 1876. *Grundzüge der Physiologie und Systematik der Sprachlaute.* Wien: Carl Gerhold's Sohn.

Carlson, R., B. Granström and G. Fant. 1970. Some studies concerning the perception of isolated vowels. *STL-QPSR* (KTH, Stockholm) 2-3/1966. 19-40.

Carmody, F. 1937. X-ray studies of speech articulation. *University of California Publications in Modern Philology* 20:187-237. Berkeley: Univ. of California Press.

Catford, J. 1974. Phonetic fieldwork. In *Current Trends in Linguistics* 12, T. Sebeok, ed. The Hague: Mouton. pp. 2489-2505.

Chiba, T. and M. Kajiyama. 1941. *The Vowel—Its Nature and Structure*. Tokyo.

Chladni, E. 1809. *Traité d'Acoustique*. Paris.

Chomsky, N. 1964. Current issues in linguistic theory. *The Structure of Language*, J. Fodor and J. Katz, eds. Englewood Cliffs, NJ: Prentice Hall. pp. 50-118.

_____ and M. Halle. 1968. *The Sound Pattern of English*. New York: Harper.

Cooley, W. and P. Lohnes. 1971. *Multivariate Data Analysis*. New York: Wiley.

Cooper, F., P. Delattre, A. Liberman, J. Borst and L. Gerstman. 1952. Some experiments on the perception of synthetic speech sounds. *JASA* 24.597-606.

Delattre, P. 1951. The physiological interpretation of sound spectrograms. *PMLA* 66.864-875.

_____, A. Liberman and F. Cooper. 1951. Voyelles synthétiques à deux formantes et voyelles cardinales. *Le Maître Phonétique* 96.30-36.

Du Bois-Reymond, F. 1812. *Fragmente aus Kadmus*. Die Musen. Berlin.

Edwards, A. 1972. *Likelihood: An Account of the Statistical Concept of Likelihood and Its Applications to Scientific Inference*. Cambridge: Cambridge Univ. Press.

Fant, G. 1959. Acoustic analysis and synthesis of speech with application to Swedish. *Ericsson Technics* 15.3-108.

_____. 1960. *The Acoustic Theory of Speech Production*. The Hague: Mouton.

_____. 1966. A note on vocal tract size factors and non-uniform F-pattern scaling. *STL-QPSR* (*KTH*, Stockholm) 4/1966.22-30.

Fant, G. 1975. Non-uniform vowel normalization. *STL-QPSR* (*KTH*, Stockholm) 2-3/1975.1-19.

Flanagan, J. 1957. Estimation of the maximum precision necessary in quantizing certain "dimensions" of vowel sounds. *JASA* 29.533-534.

Frøkjær-Jensen, B. 1967. Statistical calculations of formant data. *Annual Report of the Institute of Phonetics of the Univ. of Copenhagen* 2.158-170.

Gay, T. 1974. A cinefluorographic study of vowel production. *Journal of Phonetics* 2.255-266.

_____ and T. Ushijima. 1974. Effect of speaking rate on stop consonant-vowel coarticulation. *Preprints of Speech Communication Seminars*, Vol. 1.205-208. Stockholm: KTH.

_____, _____, H. Hirose and F. Cooper. 1974. Effects of speaking rate on labial consonant-vowel coarticulation. *Journal of Phonetics* 2.47-63.

Gerstman, L. 1968. Classification of self-normalized vowels. *IEEE Trans. Audio Electro. Acoust.* 16.78-80.

Halle, M. 1964. On the bases of phonology. *The Structure of Language*, J. Fodor and J. Katz, eds., Englewood Cliffs, NJ: Prentice Hall. pp. 324-333.

Hanson, G. 1967. Dimensions in speech sound perception. *Ericsson Technics* 23.3-75.

Harris, K. 1974. Physiological aspects of articulatory behavior. In *Current Trends in Linguistics* 12: phonetics, T. Sebeok, ed., The Hague: Mouton. pp. 2281-2302.

Hellwag, C. 1781. *De Formatione Loquelae*. Dissertation, Tübingen.

Holder, W. 1967. *Elements of Speech*. Leeds: Scholar Press.

Houde, R. 1967. *A Study of Tongue Body Motion During Selected Speech Sounds*. Dissertation, University of Michigan.

Hurford, J. 1969. The judgment of vowel quality. *Language and Speech* 12.389-395.

International Phonetic Association. 1949. *Principles of the International Phonetic Association*. London.

Jakobson, R, G. Fant and M. Halle. 1952. Preliminaries to speech analysis. *Technical Report* 13, MIT Acoustics Laboratory. 5th printing, 1963. Cambridge: MIT Press.

Jones, D. 1928. *Das System der Association Phonetique Internationale (Weltlautschriftverein)*. Berlin.

_____. 1969. *An Outline of English Phonetics*. Cambridge: W. Heffer & Sons.

Joos, M. 1948. Acoustic phonetics. *Lg.* 24 (supplement).

Koopmans-van Beinum, F. Comparative phonetic vowel analysis. *Journal of Phonetics* 1.249-261.

Kuehn, D. 1973. *A Cinefluorographic Investigation of Articulatory Velocities*. Dissertation, University of Iowa.

Ladefoged, P. 1960. The value of phonetic statements. *Lg.* 30.387-396.

_____. 1967. *Three Areas of Experimental Phonetics*. London: Oxford Univ. Press.

_____. 1971a. *Preliminaries to Linguistic Phonetics*. Chicago: Univ. of Chicago Press.

_____. 1971b. The limits of phonology. *Form and Substance*, B. Spang-Thomsen, ed., Copenhagen: Akademisk Forlag. pp. 47-56.

_____ and D. Broadbent. 1957. Information conveyed by vowels. *JASA* 29.98-104.

_____, J. DeClerk, M. Lindau and G. Papcun. 1972. An auditory-motor theory of speech production. *UCLA Working Papers in Phonetics* 22.48-75.

Laver, J. 1965. Variability in vowel perception. *Language and Speech* 8.95-121.

Lazicius, G. 1961. *Lehrbuch der Phonetik*. Berlin: Akademic-verlag.

Liberman, A., F. Cooper, K. Harris and P. MacNeilage. 1962. A motor theory of speech perception. *Proc. Speech Communication Seminars*, Vol. 2. Stockholm: Royal Institute of Technology.

_____, _____, K. Harris, P. MacNeilage and M. Studdert-Kennedy. 1967. Some observations on a model for speech perception. In *Models for the Perception of Speech and Visual Form*, W. Wathen-Dunn, ed., Cambridge, MA.: MIT Press.

Liberman, A., F. Cooper, D. Shankweiler and M. Studdert-Kennedy. Perception of the speech code. *Psych. Review* 74.431-461.

Lieberman, P. 1970. Towards a unified phonetic theory. *Ling. Inq.* 1.307-322.

_____. 1976. Phonetic features and physiology: A reappraisal. *Journal of Phonetics* 4.91-112.

Lindblom, B. 1963. Spectrographic study of vowel reduction. *JASA* 35.1773-1778.

_____. 1964. Articulatory activity in vowels. *STL-QPSR* (*KTH*, Stockholm) 2-3/1964.1-5.

_____ and B. Sundberg. 1969. A quantitative model of vowel production and the distinctive features of Swedish vowels. *STL-QPSR* (*KTH*, Stockholm) 1/1969.14-32.

_____, _____. 1971. Acoustic consequences of lip, tongue, jaw and larynx movement. *JASA* 50.1166-1179.

Lobanov, B. 1971. Classification of Russian vowels spoken by different speakers. *JASA* 49.606-608.

Lotz, J., A. Abramson, L. Gerstman, F. Ingemann and W. Nemser. 1960. The perception of English stops by speakers of English, Spanish, Hungarian and Thai: A tape cutting experiment. *Language and Speech* 3.71-77.

MacNeilage, P. 1970. Motor control of serial ordering of speech. *Psych. Review* 77.182-196.

Mattingly, I. and A. Liberman. 1969. The speech code and the physiology of language. In *Information Processing in the Nervous System*, K. Liebovic, ed., Berlin: Springer-Verlag. pp. 97-117.

Mermelstein, P. 1973. Articulatory model of speech production. *JASA* 53.1070-1082.

Michaelis, G. 1881. *Über die Anordnung der Vokale*. Berlin: von Barthol.

Miller, R. 1953. Auditory tests with synthetic vowels. *JASA* 25.114-121.

Nie, N., C. Hull, J. Jenkins, K. Steinbrenner and D. Brent. 1975. *Statistical Package for the Social Sciences* (*SPSS*). New York: McGraw-Hill.

Nordström, P. 1975. Attempts to simulate female and infant vocal tracts from male area functions. *STL-QPSR* (*KTH*, Stockholm) 2-3/1975.20-33.

_____ and B. Lindblom. (forthcoming). A normalization procedure for vowel formant data. *Proc. Eighth Intl. Cong. of Phonetic Sciences in Leeds 1975.*

Parmenter, C. and S. Treviño. 1932. Vowel positions as shown by X-ray. *Quart. J. of Speech* 18.351-369.

Passy, P. 1888. Kurtze Darstellung des französischen Lautsystems. *Phonetische Studien* 1.18-40.

_____. 1890. *Etude sur les changements phonétiques et leurs caractères généraux.* Paris.

Perkell, J. 1965. *The Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study.* Cambridge, MA: MIT Press.

_____. 1971. Physiology of speech production: A preliminary study of two suggested revisions of the features specifying vowels. *Quarterly Progress Report*, MIT Research Laboratory of Electronics, 102.123-139.

Peterson, G. 1951. The phonetic value of vowels. *Lg.* 27.541-553.

_____. 1961. Parameters of vowel quality. *JSHR* 4.10-29.

_____ and H. Barney. 1952. Control methods used in a study of vowels. *JASA* 42.175-184.

Pike, K. 1943. *Phonetics.* Ann Arbor: Univ. of Michigan Press.

Pols, L., H. Tromp and R. Plomp. 1972. Frequency analysis of Dutch vowels from 50 male speakers. *JASA* 53.1093-1101.

_____, L. van der Kamp and R. Plomp. 1969. Perceptual and physical space of vowel sounds. *JASA* 46:458-467.

Potter, R. and J. Steinberg. 1950. Toward the specification of speech. *JASA* 22.807-820.

Raphael, L. and F. Bell-Berti. 1975. Tongue musculature and the feature of tension in English vowels. *Phonetica* 32.61-73.

Ringgard, K. 1965. The phonemes of a dialectal area perceived by phoneticians and by the speakers themselves. *Proc. Fifth Intl. Cong. of Phonetic Sciences in Munster* 1964.495-501. New York: S. Karger.

Russell, G. 1928. *The Vowel*. Columbus: Ohio State Univ. Press.

Singh, S. 1974. A step towards a theory of speech perception. *Preprints of the Speech Communication Seminar* 1974, vol. 3.55-66. Stockholm: KTH.

_____ and D. Woods. 1971. Perceptual structure of 12 American English vowels. *JASA* 49.1861-1866.

Stevens, K. and A. House. 1955. Development of a quantitative description of vowel articulation. *JASA* 27.484-493.

_____. 1961. An acoustical theory of vowel production and some of its implications. *JSHR* 4.303-320.

_____. 1963. Perturbation of vowel articulations by consonantal contexts: An acoustical study. *JSHR* 6.111-128.

Stevens, S. and J. Volkmann. 1940. The relation of pitch to frequency: A revised scale. *Amer. J. Psychol.* 53.329-353.

Strange, W., R. Verbrugge, D. Shankweiler and T. Edman. 1976. Consonant environment specifies vowel identity. *JASA* 60.213-224.

Studdert-Kennedy, M. 1974. The perception of speech. In *Current Trends in Linguistics*, vol. 12: phonetics. T. Sebeok, ed., The Hague: Mouton. pp. 2349-2385.

Sweet, H. 1910. Phonetics. *Encyclopedia Britannica*, 11th ed. Vol. 21.458-467. New York.

_____. 1971. *The Indespensible Foundation: A Selection from the Writings of Henry Sweet*. Ed. by E. Henderson. London: Oxford University Press.

Tatsuoka, M. 1971. *Multivariate Analysis: Techniques for Educational and Psychological Research*. New York: Wiley.

Tatham, M. 1971. Classifying allophones. *Language and Speech* 14.140-145.

Terbeek, D. and R. Harshman. 1971. Cross language differences in the perception of natural vowel sounds. *UCLA Working Papers in Phonetics* 19.26-38.

Thompson, C. and H. Hollien. 1970. Some contextual effects on the perception of synthetic vowels. *Language and Speech* 13.1-13.

Trubetzkoy, N. 1969. *Principles of Phonology*. Tr. by A. Baltaxe. Berkeley: University of California Press.

Van der Stelt, J., J. Blom and L. Van Herpt. 1973. Stability of vowel systems. *Proc. Institute of Phonetic Sciences*, Univ. of Amsterdam 3.33-41.

Verbrugge, R., W. Strange, D. Shankweiler and T. Edman. 1976. What information enables a listener to map a talker's vowel space? *JASA* 60.198-212.

Vietor, W. 1898. *Elemente der Phonetik des Deutschen, Englischen und Französischen*. Leipzig: Reisland.

Wallis, J. 1972. *Grammar of the English Language*. Translation and commentary by J. A. Kemp. London: Longman.

Wheatstone, C. 1879. *The Scientific Papers of Sir Charles Wheatstone*. London: Taylor and Francis.

Wilkins, J. 1668. An essay towards a rean character and a philosophical language. *IBM Microbook LEL* 11135.

Winer, B. 1971. *Statistical Principles in Experimental Design*. New York: McGraw-Hill.

Witting, C. 1962. On the auditory phonetics of connected speech: Errors and attitudes in listening. *Word* 18.221-248.

Zwirner, E. and K. Zwirner. 1970. Principles of phonometrics. Trans. by H. Bluhme. *Alabama Linguistics and Philological Series* no. 18. University AB: University of Alabama Press.