A representation for the information-carrying units of natural speech
by William Phillips Rupert

A thesis submitted to the Graduate Faculty in partial fulfillment of the requirements for the degree of DOCTOR OF PHILOSOPHY in Electrical Engineering
Montana State University

Abstract:
A new representation of human speech forms the basis for the design of a system for extracting the linguistically encoded information of natural speech. We restrict linguistically encoded information to the words of an utterance and require that any representation preserve sufficient information to allow the emulation of major aspects of human listener behavior. The parameterization of the acoustic speech signal begins with a predictive segmentation that provides boundaries of signal epochs of homogeneous character. The acoustic details of each segment are recorded, and the segments are classified as one of eleven Production Modes. A Relative Opposition characterizes the distinctive changes in the signal character between adjacent PM's. The ROi is an ordere'd set of elementary calculations that provide a detailed description of the relationship between PMi and PMi+1. . The complete representation of an utterance is then a sequence of elements (PM's) and the relationships between them (RO's) that are consistent over divers speakers.

The relevance of the representation is discussed with respect to a general stratificational linguistic theory model of the speech encoding process. The PM-RO sequences are decoding units in a P-PHON stratum that describes the interface between the phonetic elements of the PHON stratum and the acoustic speech signal waveform. The use of these units in the design of an automated decoding system will realize two very important attributes.

1. The emulation of five general characteristics of human listener behavior: a. Operation on the conversational speech of divers speakers.

b. Potential real-time operation.

c. Strong immunity to noise and interference.

d. Incremental changes in the vocabulary.

e. Probabilistic operation on several levels to provide recognition, perception, error detection and correction, and nonsense response.

2. A significant economy of computation resulting from the directed search induced by the representation.

A REPRESENTATION FOR THE INFORMATION-CARRYING UNITS OF NATURAL SPEECH

by

WILLIAM PHILLIPS RUPERT, JR.

A thesis submitted to the Graduate Faculty in partial fulfillment of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

in

Electrical Engineering

Approved:

Head, Major Department

Chairman, Examining Committee

Graduate Dean

MONTANA STATE UNIVERSITY
Bozeman, Montana

August, 1969

ACKNOWLEDGEMENT

The author wishes to express his sincere appreciation to Dr. W. H. Foy and Professor N. A. Shyne for their assistance and guidance in this research and their aid in the preparation of this manuscript. Thanks are also extended to E. J. Craighill and other staff members of Stanford Research Institute for the many hours of critical discussion concerning this manuscript.

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

A new representation of human speech forms the basis for the design of a system for extracting the linguistically encoded information of natural speech. We restrict linguistically encoded information to the words of an utterance and require that any representation preserve sufficient information to allow the emulation of major aspects of human listener behavior. The parameterization of the acoustic speech signal begins with a predictive segmentation that provides boundaries of signal epochs of homogeneous character. The acoustic details of each segment are recorded, and the segments are classified as one of eleven Production Modes. A Relative Opposition characterizes the distinctive changes in the signal character between adjacent PM's. The RO_i is an ordered set of elementary calculations that provide a detailed description of the relationship between PM_i and PM_{i+1}. The complete representation of an utterance is then a sequence of elements (PM's) and the relationships between them (RO's) that are consistent over divers speakers.

The relevance of the representation is discussed with respect to a general stratificational linguistic theory model of the speech encoding process. The PM-RO sequences are decoding units in a P-PHON stratum that describes the interface between the phonetic elements of the PHON stratum and the acoustic speech signal waveform. The use of these units in the design of an automated decoding system will realize two very important attributes.

1. The emulation of five general characteristics of human listener behavior:
   a. Operation on the conversational speech of divers speakers.
   b. Potential real-time operation.
   c. Strong immunity to noise and interference.
   d. Incremental changes in the vocabulary.
   e. Probabilistic operation on several levels to provide recognition, perception, error detection and correction, and nonsense response.
2. A significant economy of computation resulting from the directed search induced by the representation.

1.0 INTRODUCTION

For many years, activity in numerous research areas has aimed toward a fuller description and a deeper understanding of man's verbal communication processes. The recent advent and availability of electronic computing equipment has accelerated the collection of data and detailed analysis concerning many facets of speech production and analysis. The possibility of man-machine dialogue in man's natural language has motivated considerable work in the general area of automatic speech recognition. Specific automatic recognition studies have generally been directed at either the message content, the speaker's identity, or the speaker's emotional state. In this study, we are concerned with efficient methods for machine decoding of the linguistically-encoded information embedded in human speech.

The foremost problem in machine recognition, regardless of the specific objectives, is that of partitioning the continuous speech signal into time epochs (segments of similar acoustic character) common in both form and function to the speech of all members of a given language community. In the case of the linguistic content, a subjective solution is available in terms of the minimum sound units that listeners identify. The assignment of symbols to these minimum detectable sound units and the description of their significance within a particular language is the branch of linquistics known as phonetics. The phonology of a particular language is the grouping together of the phonetic transcriptions of that language's minimum sound units according to significance with respect to semantic content. The linguistic system of a language includes a phonology and a

grammar. The phonology identifies significant sounds with respect to meaning; the grammar contains the rules for the expression of ideas. The elements of a phonology are called phonemes. The substitution of one phoneme for another changes the identity of the utterance. The elements of the grammar are words, phrases, and groups of phrases; the rules of the grammar specify constructions that produce well-formed utterances.

An objective solution must deal only with measurements on the physical speech signal as opposed to the subjective properties ascribed to it by human listeners. Different speakers, due to their personal physiology, produce different acoustic waveforms. Isolated speech sounds have no absolute meaning in themselves; they have meaning only with respect to other sounds. If we view speech as the concatenation of phonemes, then the realization of each phoneme is modified by its adjacent phonomes. What is even more disturbing is that the boundaries of adjacent phonomes blend together to form a continum of acoustic energy. Many of the acoustic cues that appear significant in the human identification of sounds seem to be embedded in the transitional portions of the speech waveform. In any recognition system, the processing of these transitory portions of the signal is extremely important and is recognized as a major factor in the potential performance of a total recognition system [Flanagan][1].

Several automatic speech recognition systems have achieved a significant level of performance operating with a high-quality signal from a small class of speakers uttering a limited set of words and phrases. These systems are engineering solutions to well-specified problems that are susceptible, by their very nature, to the application of state-of-the-art

pattern recognition techniques. In the area of speech production, using the information gained from speech analysis, several classes of synthesizers have been used to construct words and phrases that readily convey information. However, listensers invariably attribute an unnatural mechanical sound to the synthesized speech. These efforts in automatic analysis and production have generated a considerable store of knowledge about the particulars of the physical speech signal and the redundancy of acoustic cues resident to it. But no comprehensive theory has resulted.

The basic sound units used in the current investigations have not proved suitable as basic units of a consistent theory of speech. All current investigators have chosen either to attempt to imitate the human phonemic analysis or simply to define convenient sound units in terms of their analysis equipment and the problem at hand. In this paper, we attempt to define an information-carrying unit from the physical unit that possesses the latent power to serve as the basis for a system able to emulate many of the facets of the human listener's behavior. The elements of a phonology constructed from these physical sound units might be called "p-phonemes." The next section contains a brief description of the sound elements of spoken language in pseudo-technical classical terms. Although more qualitative than quantitative, it provides a common framework within which to discuss the higher level aspects of spoken communication. In Section 4.0, we will re-open a quantitative discussion of the physical elements of speech.

1.1 A BRIEF DESCRIPTION OF SPEECH ELEMENTS

The human production of speech consists of a series of physical movements controlled by patterned nervous impulses and performed in sequence to produce audible agitation of the surrounding air. The source of energy is a steady stream of air exhaled from the lungs. The air flows outward through the vocal tract, various parts of which create acoustical disturbances perceived as sounds. The larynx, situated just above the lungs, contains the vocal cords. The area above the larynx is considered the vocal tract; it consists of the pharynx, the mouth and the nose. The acoustical shape of the vocal tract is varied by movements of the tongue, lips, and muscles of the throat and jaw. The process of adjusting the vocal tract characteristics to produce different speech sounds is called articulation.

The vocal cords form an adjustable barrier for the air flow from the lungs into the vocal tract. When they are relaxed and open, air may pass freely into the vocal tract. If the vocal tract is constricted at one or more points, the air stream becomes turbulent and produces a hiss or fricative-like sound (e.g. /f/ or /s/). If the vocal cords are tensed to the point of periodically interrupting the air flow, then a periodic train of air puffs is produced. This periodic train of air puffs is heard as a buzz sound, and after passing the length of the vocal tract is recognized as the pitch of a person's voice. The buzz source produces a distribution of energy at harmonics of the basic pitch frequency. This spectrum of energy is shaped by the other elements of the vocal tract to produce sounds heard as vowels or vowel-like. A sound may consist solely of voiced energy

(periodic excitation) or noise-like energy (aperiodic excitation), or it may be a combination of the voiced and noise-like excitations. Sounds are also given a unique characteristic when the soft palate opens the path in the nasal cavities, allowing acoustic energy to radiate from the nostrils. Many sounds are transitory in nature and may be identified only when produced in combination with adjacent sounds. These are the stop sounds, where there is a total constriction at some point in the vocal tract, followed by a rapid release.

The distinctive sounds within a language from which words and phrases are built are called phonemes. The phoneme does not symbolize the exact acoustic description of any sound. But phonemes have meaning in relation to other phonemes; they distinguish one word from another. Accordingly, the three forms of the English consonant /k/ in the initial position of the words "key", "car", and "cool" are identified as one phoneme, while in Slavic languages this initial sound would be identified as three different phonemes. The concept "phoneme" and a more precise definition of its character will be discussed in a later section. For the present, we shall consider phonemes to be the sounds within a particular language that the language community members identify as the building blocks of their speech.

Phonemes are generally produced in groups of two's or three's, which we know as syllables. The syllable usually has a vowel as its central phoneme, surrounded by one or more consonants ordinarily of lesser amplitude. The vowel is generally the strongest and longest portion of the syllable and is heard as being modified by the surrounding consonants. A fairly general set of English vowels is the following [Pike][2].

Pure Vowels		Examples of Diphthongs	
i	feet	ou	tone
I	fit	ei	take
e	mate	ai	might
E	said	au	shout
ae	cat	oi	toil
Λ	cup		
u	boot		
U	foot		
o	rote		
ɔ	caught		
a	colony		

Diphthongs are combinations of two vowels in which the change in the vocal tract from one vowel position to another vowel position is a significant characteristic. English also has two semi-vowels: /w/ and/y/. These sounds are formed by briefly positioning the vocal tract in a vowel-like configuration and then rapidly changing it into the form required by the vowel of the syllable. The semi-vowels may not be formed in isolation and must always precede a vowel. One may regard /w/ and /y/ either as vowels or as consonants.

The consonants of English may be grouped for discussion according to their manner of articulation, place of articulation, or their distribution within continuous utterances. We have grouped them by manner of articulation and will discuss place of articulation within these categories. No discussion of distributional properties seems appropriate at this point.

Nasals (m, n, ŋ)

The nasal sounds are formed when the soft palate is opened to allow the nasal cavity to become a second branch of the vocal tract. They are voiced sounds, continuous in character throughout their duration. The immobility of the nasal cavity fixes the distribution of acoustic energy throughout the nasal sounds. Articulation may take place anywhere from the front of the mouth to the back, the /m/ being formed with the lips and the /ŋ/ being formed in the back of the mouth.

Liquids (l, r)

The liquids are articulated in the interior of the mouth. Their duration is long, like that of a vowel. The /l/ is unique in that the tip of the tongue is placed against the roof of the mouth and an acoustic cavity is formed on either side of it. The /r/ sound is always transitory; in an initial position it involves the movement of the tongue tip from the roof of the mouth to the following vowel position, while in the terminal position the vowel ends with the tip of the tongue moving to the roof of the mouth. The /l/ and /r/ are both voiced sounds with an amplitude slightly less than that of the pure vowels.

Fricatives (f, v; θ, th; s, z; sh, zh; h)

The fricative sounds in English have as a significant feature a wide spectral distribution of noise-like energy. The fricative sounds come in pairs, grouped according to their place of articulation. The first sound in the list, /f/, is pure noise-like energy, while its voiced cognate, /v/, has the same noise-like energy with the addition of low-frequency voicing.

The fricative sounds are continuous and may be produced in isolation. Their duration approaches the length of vowels, nasals, and liquids. The place of articulation progresses rearward from the front of the mouth. The /f/ and /v/ are produced with the lips, while the /h/ is produced in the pharynx.

Stops (p, b; t, d; k, g)

All stop conconants depend upon the dynamic action of the vocal tract for their identity. These sounds are produced with a complete closure at some point in the vocal tract. Pressure is built up behind this occlusion, and its sudden release is characterized by an abrupt motion by the articulator. The sudden release and aspiration of turbulent air creates a very short noise-like pulse. This noise pulse may take place either with or without simultaneous voicing. In the preceding list /p/, /t/, and /k/ are the voiceless stops. Their voiced cognates are /b/, /d/, and /g/. None of them may exist or have meaning in isolation. Their place of articulation varies from the lips through the tip of the tongue to the soft palate closure. It is known that much of the information identifying stop consonants is contained in the surrounding vocalic segments.

Affricates (tf as in chew, tzh as in jar)

The affricates are analogous to the diphthongs. They are the combinations of two fricatives produced with a significance attached to the vocal tract change from the first to the second fricative.

An information-carrying aspect of speech not yet discussed is the stress and intonation of an utterance. These are the parts of the spoken language that express a speaker's emotional attitude and semantic goals. By the use of stress and intonation, one may drastically alter the semantic content of an utterance. Stress and intonation are generally directed at syllables and thus affect the speech utterance over a longer time unit than the phoneme. The major influence in a stressed syllable is on the vowel. Three factors together indicate stress: amplitude, increase in duration, and change in pitch. These factors fit together within the overall intonation of the utterance.

It must be emphasized that no particular set of rules exists for associating an acoustic speech waveform segment unambiguously with one of the above phoneme labels. The phoneme is the name for a set of articulatory movements having meaning only when compared to some other set. Furthermore, their distinctiveness can be observed only as they function to differentiate one word from another. The concatenation of phonemes in the syllable forces some change in acoustic character of each individual phoneme. In addition, the concatenation of syllables causes the boundary phonemes (between syllables) to undergo further modification.

The electrical signal as transduced by an acoustical-to-electrical device has some rather complicated properties. The passband of high-quality speech is approximately 80 cycles per second to 10 kilocycles per second. Normal telephone-quality speech used in everyday communication is bandlimited between 300 cycles per second and 3,300 cycles per second. The particular acoustic segments that compose phonemes vary in duration from

about 20-millisecond stop sounds to 500-millisecond vowel sounds. The amplitude variation in a normal connected utterance will be as great as 20 to 30 db. The vowel-like fricative sounds are very low in amplitude and peak-to-average ratio. In the spectrum of speech signals we observe that the energy of most sounds is concentrated at one or more frequencies. These regions of the spectrum are denoted as formants and generally numbered from 1 to 4, starting at the low-frequency end. The vowel sounds are normally characterized by three formant positions within this range. The noise-like portions of the fricative and plosive sounds may extend over a 2- or 3-kc region. Their energy may be concentrated in a single peak or in two peaks, with a possible third and lowest formant (that is, the voicing). Brief silences within speech surround the stop sounds.

Figure 1 is a sonagram of an isolated spoken word. The sonagram is made on a Kay Sonagraph machine. As indicated on the figure, the horizontal axis displays time at 200 milliseconds per inch, and the vertical axis is a linear frequency scale. The voiceless fricative (at A) is indicated along with the formants. The darkness of the shading represents the amount of energy.

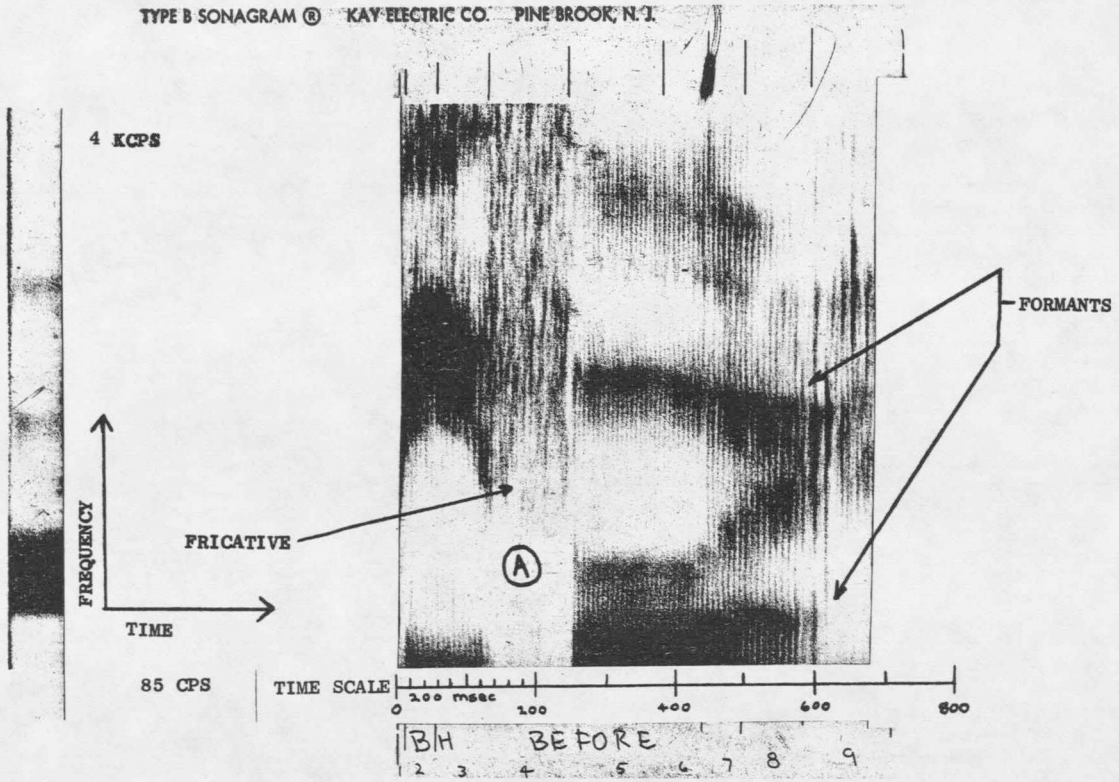


Figure 1. Sample Sonagram of /Before/.

1.2 CONTEXT AND INFORMATION DISTRIBUTED IN NATURAL SPEECH

Human speech and the phenomenon of spoken communication is a complex process requiring the participation of both the speaker and the listener. To describe in detail all the information contained in a particular spoken utterance is a difficult, if not impossible, problem. One might as well be trying to define the beauty in a painting. A large measure of the meaning conveyed by a spoken segment is in the ears of the beholder. More formally, a good deal of the information transmitted in speech depends upon the listener's knowledge of the subject matter, the speaker's identity, and the speaker's motivation. For the sake of discussion, we will define three primary types of information contained in any complete utterance. These three classes of information are in no way disjoint, nor is some information from each class required to be present for the existence of an utterance. Of the three types of information listed, we are primarily concerned with methods of automatically extracting the linguistically encoded information from natural speech.

1. Linguistically Encoded Information -- is considered to be the equivalent orthographic representation of the words of the utterance. The determination of the words in the utterance should be the consensus of a group of ordinary English speakers.

2. Speaker Identity Information -- is considered to be that portion of the speech signal primarily determined by the speaker's physiology and his manner of articulation. A secondary, long-term identification of the speaker's identity is certainly provided by his choice of words.

3. Speaker's Emotional State Information -- is the information that tells listeners familiar with the speaker of the motivation for the utterance. Three facets of the speaker's emotional state seem appropriate.

a) The speaker's quiescent state of mind;

b) The state of mind which the speaker wishes to present to his listener;

c) Any a priori reaction induced in the speaker by the reaction he expects to elicit from his current utterance.

To complicate matters further, spoken communication is arranged so that there are at least three levels of context within which the information is embedded. In one sense, a level of context is really no more than an acceptable format for presenting a particular type of information. Three levels of context with fuzzy boundaries between them may be outlined.

1. Global Context -- This most inclusive level is a combination of the speaker's semantic intention and the rapport between the speaker and the listener. This level is often considered to be the speaker's tone of voice.

2. Local Context -- This is the grammatical level that specifies the word arrangement required to transmit the linguistically encoded content.

3. Micro-Context -- This level is also known as the phonotactics of a language. It is the collection of acceptable modifications which the speaker may incorporate in the formation of the individual units of the sequence he is articulating. Most often these micro-context features are not noted separately by the speaker or the listener.

These levels of context operate in an hierarchical fashion directed downward from global to micro. However, the time span of the context dictated at each level is much greater than the span of the level below it. Thus, the context which sets the tone of voice does not specify the choice of words. Nor does the choice of words dictate the articulatory stresses to be used in their production. The individual sound articulation is affected by the global context.

The most fortuitous situation would be to have each of the three types of information appearing in a separate level of context. However, this is not the case. The extraction of the linguistically encoded content of an utterance is aided by the global and local context. In fact, many times the listener relies on what he expects to hear to help resolve ambiguities arising at the micro-context level. The amount of linguistically encoded information, the number of words correctly understood in an utterance, is the only concrete thing which lends itself to easy measurement.

The amount of information contained in a speaker's signature is a nebulous quantity. His personal signature invades all levels of the context. That portion of his signature due to his physiology shows up in the

micro-context as the range of his voice, the strength of his voice, and his speed of articulation. His personal signature also shows up in his choice of words and his tone of voice at the global context level. At the intermediate context level, his dialectal background colors his pronunciation. The amount of this type of information transferred in any particular utterance is unquestionably a function of the listener's familiarity with the speaker and the situation. The information about the speaker's emotional state shows up at all levels of context. However, this type of information may be evaluated only with respect to some quiescent state of the speaker. In one sense, the speaker's deviations from his quiescent emotional state become his emotional signature. It is apparent that no determination of the speaker's emotional state could be reliably undertaken with one or two isolated samples of his speech. Any attempt to assign a quantitative measure on the amount of information concerning the speaker's identity and emotional state contained in a particular utterance is senseless. An automaton designed to extract this information could not be evaluated in terms of yes or no performance. It could be said that such an automaton was successful only if it aided in determining the linguistic content of utterances.

The remainder of this paper will be concerned with the extraction of the linguistically-encoded information from a spoken message. The problem is that of converting a normal, conversational, English-language utterance to an orthographic representation with the same linguistic content. The first restriction is to accept the fact that homophonorous words (e.g. The

red car; He read the book) are the same when spoken and homographic words (e.g. Row the boat; They had a row) are different words in spoken language. The goal of the decoding process then becomes relating acoustic segments of the speech signal to meaningful groups of symbols.

1.3 LINGUISTIC SYSTEM MODEL FOR SEMANTIC CONTENT

In this section, we will detail a linguistic model deceptively similar to those proposed by Lamb [3] and other stratificational grammarians. These models are most detailed in terms of the generation of speech. They illuminate the various aspects of the speech signal generation that must be recognized and identified in decoding the utterance. The strata of the model (Figure 2) are produced in an attempt to define levels of context strictly within the semantic system. They are arranged from top to bottom by their descending or decreasing context influence in the time span sense. Another characterization of this is that the sizes of chunks of information within each stratum are comparable and locally context free. The four strata, LEX, MORPH, PHON, and P-PHON, correspond to the increasing detail that the encoding process must handle in each stratum.

Each of the strata is divided into a left and right side (with respect to Figure 2) respectively labelled with the suffixes "eme" and "on". In all strata the -eme's may be thought of as the units of significance and the -on's as the objective descriptions of these units. Through composition rules on the -eme units are translated into -on units. The transition from the -on's of one stratum to the -eme's of the next stratum is made through a set of realization rules. The -on units of a particular stratum become the corresponding -eme units of the next lower stratum through the connecting realization rules. The relationship of the -on units to their corresponding -eme units within a stratum is that of measurements to their meaningfulness. Essentially, the -on units record physical happenings, and the -eme's provide the significance of the occurrences. Consequently the

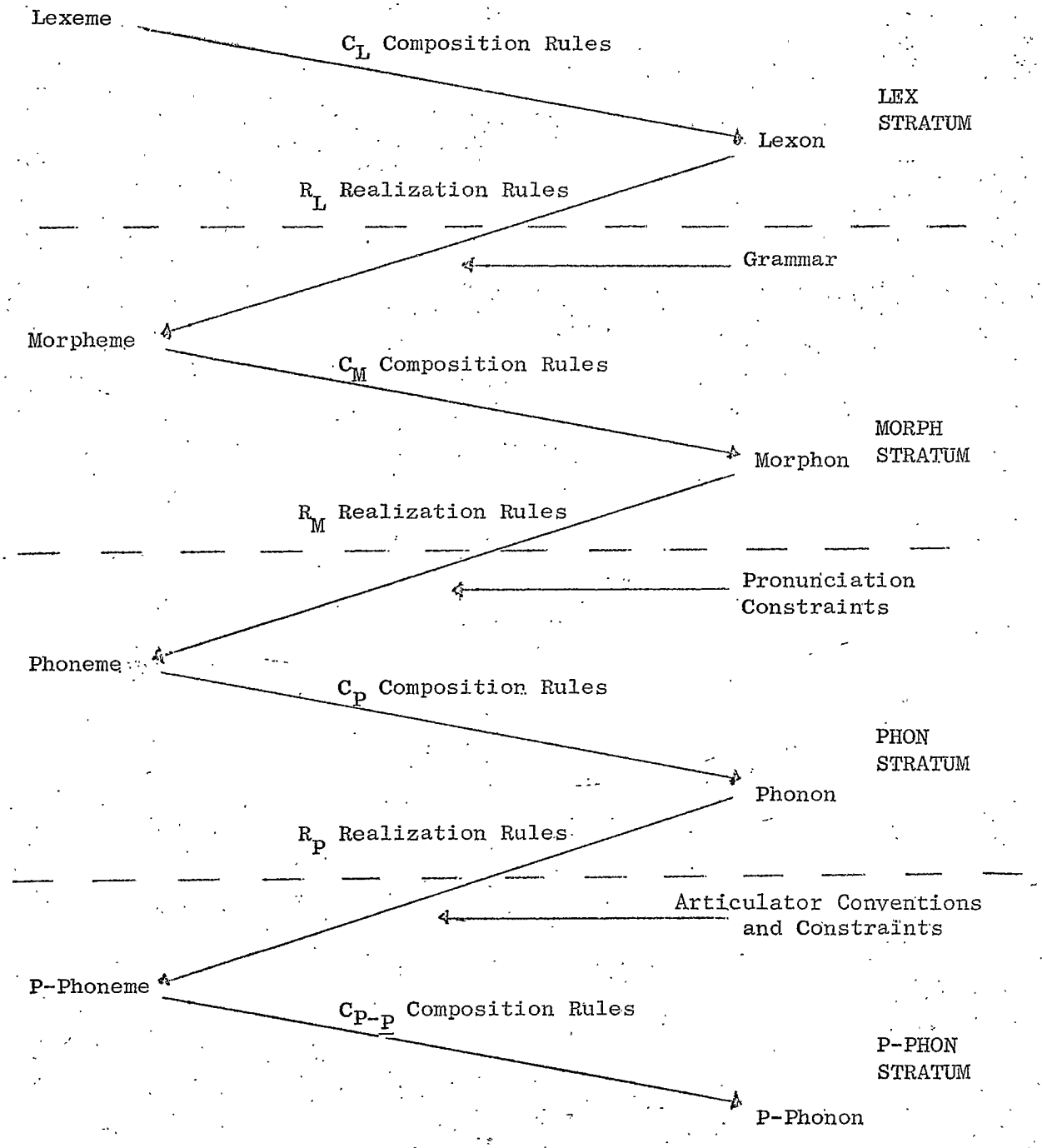


Figure 2. Structural Levels of Spoken Language as Strata

-eme's of a language may be determined only with respect to their use. Note the similarity here to the discussion of the phoneme as standing for a concept having meaning with respect to the other phonemes but not by itself.

The LEX stratum is composed of lexemes and lexons. The definition of lexeme will have to be the vague concept of a minimal unit with which the content of ideas is expressed. Lexemes are composed of lexons. The realization of lexons are the morphemes of the next lower stratum. The morpheme is quite nearly a syllable, but is more easily defined. A morpheme is a minimal unit of meaning from which words and phrases are constructed. The concept of "minimal unit of meaning" is meant to imply that the unit may not be further decomposed and still retain a stable meaning. Within the MORPH stratum, morphemes are composed of morphons, which in turn become the phonemes of the next lower stratum. Again the phoneme label represents a class of similar sounds that are normally indistinguishable within a particular language. The phoneme might be characterized as the minimum distinguishable sound unit with respect to the identity of utterances. Phonemes are composed of phonons. Phonons are a description of the articulatory features of the sound unit.

At this point in the linguistic model we have added another stratum. This stratum is called the P-PHON to indicate its close connection with the physical speech signal. The -on rank of this stratum is the minimum acoustic signal segment seen by the analysis system. The p-phoneme units would be equivalent recurring sequences of p-phonons. The addition of the P-PHON stratum is necessitated by the replacement of the human ear-brain combination with a mechanical measuring system. Consequently, we expect

many phenomena not knowingly analyzed by the human listener to appear in the P-PHON stratum. A detailed discussion of the elements in the P-PHON stratum will be supplied in Section 5.0 that relates the analysis methods to this linguistic model.

The mechanism of the composition and realization rules should be clear from an example of their operation (Figure 3). Beginning at the lexeme, we desire to discuss a collection of married women. The LEX stratum composition rules indicate that this is the wife and that it should be plural. The realization rules leading to the next stratum inform us that the plural of wife can be formed as a suffix to the word. Thus, the representation on the morpheme level is "wife" plus a plural suffix. The morph composition rules indicate that the plural of "wife" is made by the addition of "s." The morphon is thus "wife plus s." The realization rules for the phonemic representation indicate that the f changes to a v and the s to a z, producing the phonemic representation /waivz/. The composition rules of the PHON stratum transform that phonemic sequence into the appropriate set of articulatory motions. The set of articulatory motions is realized as an acoustic signal in the P-PHON stratum. Thus we encounter them as our measurements on the acoustic signal.

The strata model deals with the translation of ideas into segments of acoustic signal. It must be emphasized that it models only the generation of the linguistically encoded information. Within each stratum it deals with units of a particular size and a particular set of context rules. The grammar of a language is contained in the LEX stratum composition rules and the LEX-to-MORPH strata realization rules. The signalling elements of the

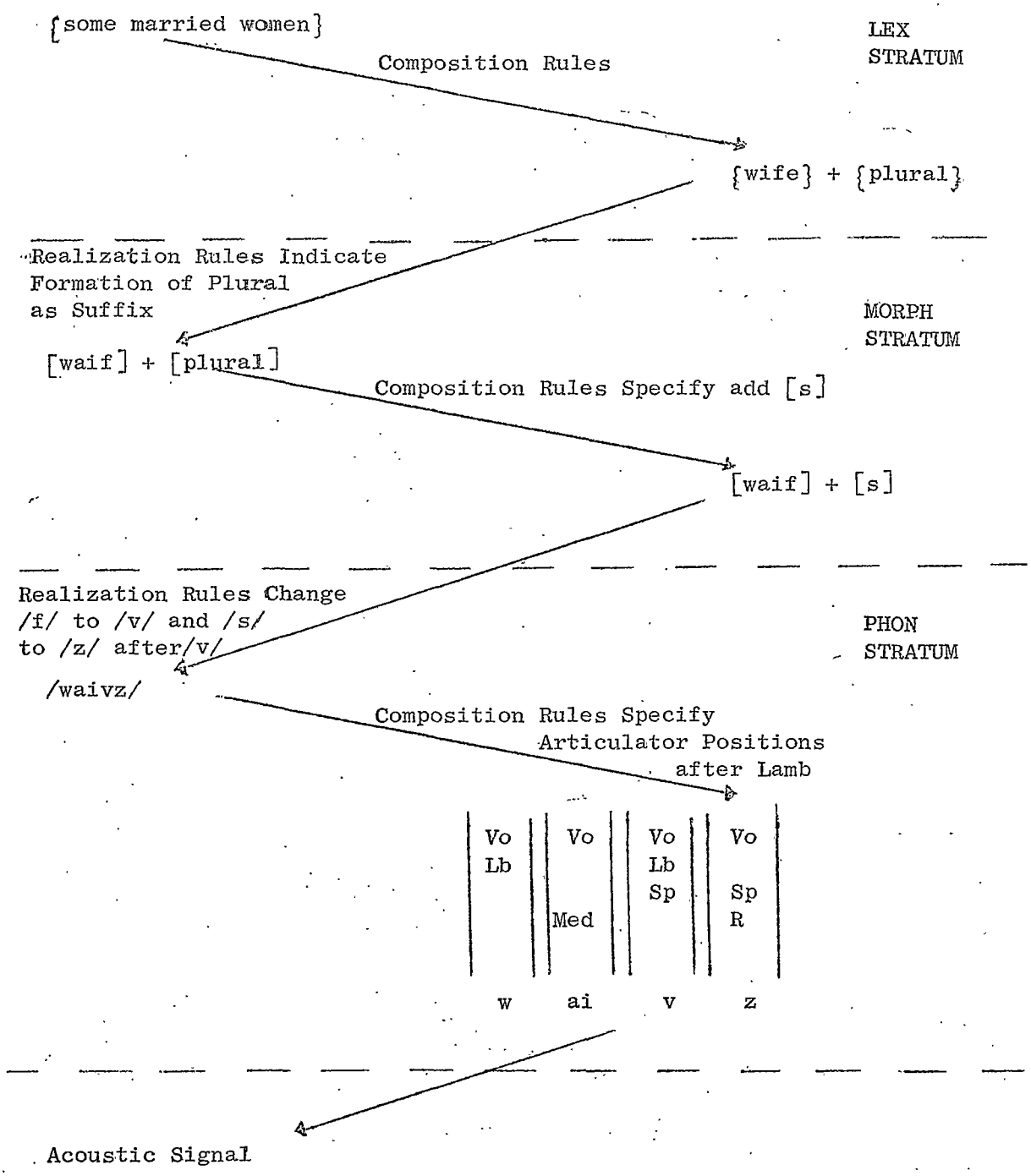


Figure 3. Example of Composition and Realization Rules

particular grammar rules are found in the acoustic signal as supra-segmental features. They modify the acoustic segments of the signal, yet the interpretation of their meaning extends over many segments. Written language attempts to realize supra-segmental features by the use of punctuation. There are easily identified supra-segmental features in the following three sentences.

Don't eat fast.

Don't eat, fast.

Don't eat fast?

The first sentence is declarative, with equal stress on each word in a monotonically descending pitch. In the second sentence, the first two words are stressed as a group, a pause in the production of the utterance is inserted at the comma, and the final word is stressed with a descending pitch. In the third sentence, the first and last words are stressed with unique pitch contours. A less prominent supra-segmental feature is the slurring of the t's in the first and third sentences. Many subtle or distributed supra-segmental features in a spoken utterance have no equivalent representation in written language. Consider the sentence:

The policeman whistled by the doorway.

This sentence may have at least two meanings.

1) The policeman may be standing by the doorway whistling.

2) The policeman may have dashed quickly past the doorway.

In spoken language, it is possible, through use of supra-segmental features, unambiguously to specify the meaning of the sentence. In written language

meaning often requires a longer length of context. The context span of the global level in written language is sometimes greater and sometimes less than the local context span in spoken language.

The lack of correspondence between the context of spoken and written language is one of the most difficult aspects to be considered in the extraction of the semantic content of an utterance. Figure 4 displays the elements of context in spoken and written language. It is obvious from this table that no correspondence exists between the two modes of expression, except that each possesses both static and dynamic context.

The strata model for the encoding of ideas into spoken language appears useful. An analogous system model operates simultaneously and in parallel to embed the speaker's identity and emotional state. This parallel system is not structured with the precision of the linguistic information encoding system. The result is that an automatic system for the extraction of the linguistic information content must operate to decode what was said without destroying the supra-segmental features specifying how it was said. This paper deals strictly with the composition of what was said.

LEVELS OF CONTEXT INFORMATION

Static Global Context -- The general relationships (physical) initially established between the message writer (speaker) and the reader (listener).

Written	Spoken
Title	Physical Situation
Author	Speaker's Identity Known from Non-Aural Means
Subject Material	Subject Material
Type of Discourse i.e. Informative, Persuasive, Rhetorical	Required-Unrequired Utterance

Dynamic Global Context

Written	Spoken
Graphical Format	The Speaker's 1) Emotional State 2) Impression He Hopes to Make 3) Any Reaction He Expects
Illustrations	Intonation (Pitch, Pitch Contours, Stress)
Tone of Writing, i.e. Formal, Colloquial, Humorous, Satirical	Timing of Articulation Movements
	Visual Observation of Speaker

Figure 4. Levels of Context Information.

Dynamic Realization

1. Macro Constructions

Written	Spoken
Words	Syllable Stress
Sentences	Clause
Paragraphs	Breath Groups
	Phrases

2. Micro Constructions

Written	Spoken
Linear Strings of Graphemes	Strings of Morphemes Constructed of Phonemes Built of Small Acoustic Segments
Type of Character Elements i.e. Script, Hand Printed Typewritten	The Linguistic Constraints are Exhibited as Patterned (Unconscious) Behavior

Figure 4 cont'd. Levels of Context Information.

2.0 BACKGROUND FROM OTHER INVESTIGATIONS

The general term "automatic speech recognition" has come to be applied to any attempt at using electronic computing equipment to extract either the linguistic content (what was said) or the speaker's identity (who said it) from an acoustic speech signal. We are primarily interested in extracting what was said, regardless of who said it. Investigation directed at the automated extraction of the linguistic content from a speech signal follow along two basic paths. The search for components or building blocks of the speech signal that could be found in all speakers' utterances is by far the older of the two paths. The newer and currently more popular approach is a total systems view of the problem. The goal is an engineering system that solves a particular set of input-response relations. A very comprehensive survey of both approaches is available in Flanagan [1] and in Lindgren [4]. Consequently, we mention only a small number here that are directly related to this investigation.

The research dealing with the components of the speech signal revolves around detailed investigations of speech produced by numerous speakers. It is motivated by the classical concept that speech is a linear string of phonemes. The search is for acoustical correlates that are the realization of each individual phoneme. Many detailed investigations of the acoustic signal were made in an effort to find aspects of the signal which changed at phoneme boundaries. The most common result is that more changes were found in the acoustic signal than there were phoneme boundaries [Tufts][5]. Some investigators feel that the differences in the acoustic signal due to the speaker's physiology may be normalized out [Thomas][6]. These investigations

have searched for methods to normalize formant ranges across numerous speakers from a minimum and maximum measurement. This method has the obvious disadvantage of requiring a sample of the speaker saying a known utterance. The notion that phoneme units would eventually be found in continuous speech predicated several investigation of isolated phoneme recognition. These investigations produced significant knowledge about pattern recognition but proved to be unproductive when applied to continuous speech [Martin][7]. A paper by Öhman [8] best demonstrates the problems of coarticulated phonemes. He considers vowel/stop-consonant/vowel utterances with the stop consonant fixed while the preceding and succeeding vowels are changed with a number of speakers providing the data. His data clearly demonstrates the relevance of the vowel transitory portions in specifying the vowel and the consonant identity. It also contradicts the concept of fixed hubs of phonemes that was suggested in the work of Potter, Kopp and Green [9] and Libereman [10]. His data provides evidence that even in precisely articulated clusters the effects of one vowel go across the stop consonant to color the other vowel.

Several investigators have produced extensive data concerning the confusion by human listeners of isolated sounds in the presence of noise and interference. The first work in this area was by Miller and Nicely[11], with other significant contributions by Singh and Black [12] and Wickelgren [13]. In all cases, the test material consists of consonant/vowel utterances, with the listeners identifying the consonant. Collectively, the results are predictable from Öhman [8] and Potter, Kopp and Green [9], the agreement being that the confusion of isolated consonants in a consonant/vowel environment is greatest in the voiceless consonants. Voiceless consonants

are seldom confused with voiced consonants. This finding agrees with data indicating that voiceless consonants have the least effect on the following vowel formants. The voiced consonants modify the formant structure of the following vowel to a greater extent and thus provide more acoustic cues. An expected result is that listeners have the most trouble with sounds not found in their native language. The confusion matrices from these experiments are most deceptive when they suggest the possibility of a closeness measure over the consonants. This possibility of a closeness measure led many investigators to feel that a consonant in context could be recognized with a second and third choice. This is not the case, because humans do not recognize consonants in a continuous utterance; rather they perceive the consonant as part of the total utterance. (An outstanding discussion of recognition versus perception is available in Sayre [14].) Any measure over consonants in a continuous utterance must be a context-sensitive measure. That is, the consonants in a particular utterance that are close to other consonants are close only in that particular sound environment. As Öhman [8] has so concretely demonstrated, the context transcends many consonants.

The other major approaches are engineering systems that respond to a certain class of speakers and provide outputs for a number of isolated words and phrases. They are most assuredly pattern recognition systems with parameterization of the input and some sort of decision structure to identify it. As such, there is no requirement that the parameters extracted from the input signal have any particular linguistic significance. The decision

structures vary in complexity with the requirements of size and type of vocabulary changes required. Numerous characteristics of total speech recognition systems may be used in classing them for discussion. Some of these characteristics are: single speaker and multiple speakers; male, child, and female speakers; vocabulary of a fixed size or arbitrary, but finite, size; the restriction to a longest word or phrase; and whether or not the response includes a second and third choice for the input utterance.

One of the earliest systems is reported by Gazdag[15]. He used the outputs of a number of filters as dimensions and then looked at simultaneously occurring states within his multiple time series. Dealing with spoken numerals, he looked for recurring sequences of machine states within his data. He used a linear threshold decision structure. His results on recognition were good, but the hardware filters on the input automatically limit the system to a certain class of male speakers.

A current systems approach by Bobrow and Klatt [16] has the particular objective of recognizing approximately fifty words and phrases (all less than two seconds in length) uttered by a very restricted class of male speakers. Their system was to be used by only one speaker at a time but had facility for rapid adjustment to one of the other speakers. They further required the addition or deletion of single vocabulary words or phrases. Their basic system uses a specially chosen set of 19 bandpass filters to provide samples of the speech spectrum. The filters are a contiguous set with each filter bandwidth related to the average information content as determined by articulation testing [French and Steinberg][17]. The envelope functions defined by the output from each bandpass filter are

simultaneously sampled every 10 milliseconds. The quantized envelopes form a set of nineteen base functions on which smaller ranged and more complicated functions are eventually computed. The more complicated functions are thresholded to -1, 0, +1, and form a time series that represents a number of features of acoustic and linguistic significance. The time element from each of the feature sequences is compressed by saving only the points of feature changes. This serves to produce short binary sequences of less than 10 elements for each two-second utterance. The collapsing of time results in a great data savings and allows a final decision structure of reduced complexity. The structure is a tree-like process realized in a list-programming language. The percentage of correct recognitions reported is significant. They were successful with several vocabularies and the required class of speakers. However, due to the bandpass filter nature of the primary input to this system, it is not adaptable to women or children. Nor is it useful for the sequences of arbitrary length that appear in natural language.

Reddy [18] and Reddy and Robinson [19] have reported on two aspects of a total system design. The former paper concerns the construction of a phoneme recognizer designed to translate continuous (up to 2 seconds in length) samples of conversational speech into phoneme strings, and the latter paper is concerned with a set of programs for parsing error-free strings of phonemes from natural language into graphemes. Two prominent features of their work are the use of processing of the real-time waveform (no filters) and the use of only one speaker. The processing programs are *ad hoc* algorithms designed for finding stable segments in the input speech

signal. One or more input segments are equated to a phoneme. At the first level, conditional probabilities of the English language have been used. Noticeably absent from the system is the determination of any characteristics of the transitory portions of long voiced sounds. Vowel identification is based on a short segment taken from the interior of the vowel segment. The justification for the work with a single speaker appears to be the feeling that a sufficiently detailed description of one speaker over a large variety of situations will provide for the extraction of at least enough acoustic cues from the speech of any speaker. The raw time series is characterized in terms of statistics of axis crossings and amplitude variations and does not make use of any orthonormalized expansions. An interesting and useful aspect of their preprocessing is that acoustic segments are initially assigned to one or more overlapping classes and the final decision is determined from context and more detailed measurements.

Reddy and Robinson [19], in attempting a phoneme-to-grapheme translation, assume an error-free string of phonemes as input. Their graphical output is determined by a dictionary of a thousand English words represented in one or more phonetic spellings. The parsing of the input phoneme string is done by a directed search of this library. Many of the constraints of natural language are included in their parsing algorithms. However, there are enough non-uniform rules on phoneme alternations in natural language so that the operation occasionally fails catastrophically. At a failure, the computer requests the operator to enter the grapheme representation of the word and the number of phonemes it contains. It then starts over in the parsing procedure. We feel that this has greatly complicated the parsing

problem by destroying supra-segmental features of stress, pause, and pitch variations in the initial preprocessing. They attempted some error correction of the input phoneme string by defining a closeness over the phonemes. Then, when the directed search of the dictionary fails, it may be started again for the next-closest phoneme. However, we pointed out before that the closeness of phonemes has previously been measured only in the sense of recognition and not as they are actually perceived in continuous speech.

To recapitulate, many accurate measurements have been made on the character of the acoustic speech signal. Much knowledge has been gained about a listener's ability to make use of the redundancy of acoustic cues. The engineering system approach to automatic speech recognition has eliminated many features of the physical signal. As a result of these studies we are left with a tremendous amount of unorganized knowledge about the characteristics of the speech signal. We know many things that can be used in the analysis of linguistically encoded content, the speaker's identity and his emotional state. This knowledge lacks an organization that details exactly what constitutes a naturally articulated speech signal, as is certainly evidenced in the numerous speech synthesizers which all have a mechanical quality about them that may be described as a human imitating a robot. In the remaining sections of this paper we detail a method, with examples, that appears to sufficiently represent the speech signals of many speakers to enable the automatic extraction of the linguistically encoded information.

3.0 THE APPROACH IN THIS INVESTIGATION

The goal of this investigation is the definition of a set of units that will allow the automatic extraction of the linguistically encoded information from natural speech. With reference to the strata model, we are attempting to define units on the P-PHON and the PHON strata. We wish to exclude consideration of any information indicative of the speaker's identity, emotional state, or the physical situation. The motivation for our approach is the fact that the search for the elusive phoneme as a definable element of the acoustic signal has failed. If we accept the notion that speakers have said the same thing when a group of listeners hear the same thing, then it certainly appears that each speaker is using an acoustic coding of the information that apparently utilizes code elements (or symbols) agreed upon by the entire group. We tacitly assumed that some minimal set of these code elements must be present in the utterances of each speaker. The major innovation in our approach is to implement the search for and the initial comparison of these elements in the P-PHON stratum.

In this investigation, we collected samples of five speakers (three males, two females) uttering the same ten, two-syllable words in a quiet environment. The primary concern has been the search for a representation that brings out similar features of each word as uttered by all speakers. The test words were chosen for the variety of sounds; consequently few sounds are present in more than one of the vocabulary words. The single overall objective is that any representation developed must lead to a decision structure that is usable on constrained natural English. Initially, neither the allowable measurements on the acoustic signal nor the form of

the decision structure were specified. This provided the freedom to alternately specify one and then the other in hopes of arriving at a powerful and efficient combination. The process was begun with the knowledge that regardless of the technique used to display the speech samples, any comparison of elements across speakers would necessarily be done simultaneously with the development of algorithms to partition the samples. In order to evaluate the potential of the possible methods of segmentation, five characteristics of human listener performance were defined. Any future system that met these five performance requirements would emulate human behavior except for the possibility of continuously decaying memory of past utterances. The five overall system operating requirements follow:

(1) Operation on the Conversational Speech of Divers Speakers

Conversational speech is not unnaturally stressed, not precisely ennunciated, and not produced with restricted tonal patterns. It consists of the normal words and phrases used in everyday communication with one's contemporaries. However, we must require some degree of conformity; the obvious substitution of sounds or words requiring a normal listener to make a conscious translation must not be allowed. The conversational aspect is also meant to imply that any recorded samples are made in a quiet environment with the speaker casually positioned with respect to a microphone. There should not be any particular restrictions on the speaker's rate of production or timing between utterances. Divers speakers is taken to mean males and females who are mature with respect to knowledge of the spoken language.

This performance requirement is interpreted primarily with respect to the lowest level of processing. The following three aspects of the acoustical signal must be taken into account.

1. Amplitude -- long-term variations over speakers and utterances of as much as 60 db with short-term variations of 20 db being normal within connected speech.

2. Frequency Range -- male and female speakers have as much as an octave variation on the low-frequency end of the speech spectrum. Some female speakers have a considerable variation in their second formant not seen in male speakers.

3. Timing -- there appears to be a pattern to the time a particular speaker dwells on the production of each sound unit. It is commonly accepted that stress is indicated by a co-relationship of duration, amplitude, and pitch. However, the duration of sounds is not consistent enough over divers talkers to be useful as the unique characteristic of particular sounds.

These three points make it clear that an acceptable segmentation technique must be independent of what is said, how it is said, and who said it. This suggests that the procedure for segmentation must mark off boundaries of time epochs whenever the character of the signal changes independently of a second level of processing that actually parameterizes and preserves a description of the signal within each time epoch.

(2) Potential for Real-Time Operation

This is a necessary property when we consider natural language and an arbitrary set of utterances with no a priori knowledge of the maximum length. That is, if we consider a finite vocabulary of English words and then ask what is the longest sentence that may be constructed, the answer is indeterminant for any significant vocabulary size. The following two properties seem sufficient to guarantee continuing real-time operation with appropriately designed hardware.

1. The preprocessing and subsequent parameterization of the incoming signal must begin with a signal onset and continue asynchronously throughout the duration of the signal. The initial parameterization should be able to reduce the storage requirements by at least a factor of 100. This property also manifests itself as a requirement for single-pass data normalization in order to allow data to be continuously passed to the decision structure.

2. The decision structure should be actively working toward a final decision and progress one step further as each unit of parameterized input is received. In fact, we would have a decision structure consider only a finite amount of past information at any one point of time while continuously attempting to convert the input speech to the highest level of coding contained in the machine. This greatly reduces the interim memory requirements.

(3) Operational Immunity to Noise and Interference

In a communication system, we generally consider noise as aperiodic signal energy mixed with the desired signal and interference as being a periodic signal in the same relation. This is not the case in human speech communication. It is known that a good deal of the human listener's ability to operate under adverse conditions is related to situational and contextual factors. However, on the sound unit level, human speech is quite cleverly constructed so as to be inherently immune to background noises. Since speech is a sequence of sound units that are alternately periodic (vowels) and aperiodic (fricatives, stops), the character of the information signal is constantly alternating. Thus, a periodic signal behind a vowel may severely interfere with a vowel's identity. However, it has no particular effect in the identification of a fricative. It is apparent that the first-stage preprocessing (including segmentation) should be able to define the speech unit boundaries in the presence of undesired signals which lack long-term speech-like qualities, and pass these segments to the decision structure in a form sufficiently detailed to preserve the character of the original speech and yet allow any necessary further processing to eliminate the residue of undesired signals. This definitely implies an adaptive segmentation procedure.

(4) Incremental Additions and Deletions to the Vocabulary

The term "vocabulary" is used in a general sense here to indicate both individual words and allowable sequences of words. The ability to expand the vocabulary is certainly one of the most useful powers of the human listener. Any finite machine will certainly have a maximum vocabulary size dictated by either the number of words that may be stored or the number of rules required to specify the allowable sequences of a few words. The fact that we require a vocabulary to be expandable from one to the maximum, N, composed of arbitrary words in an unspecified order, restricts the class of acceptable training algorithms to those that do not require repetitive presentations of the entire data base in a specific order. This desired form of operation is a natural consequence of specifying words as sequences of identifiable information-carrying units. Then, as long as a set of representation units is sufficiently complete, new words and relations between words may be added by simply presenting examples of the new sequences.

(5) Probabilistic Operation on Several Levels: Recognition, Perception, Error Detection and Correction, and Nonsense Response

This operational characteristic necessitates a distinction between a recognition process and a perceptual process [Sayre][14]. The basic process of associating one unknown input with one of a

number of stored examples is recognition, that is, the choice of that sample from the library most nearly resembling the input. A closeness measure is generally computed to effect the comparison of the input with the library members. The likeliest library member is identified as a first choice, and the others may be denoted as second, third, etc., choices. Another manner of stating this is to require that the input elements that are close should also be close in the library. We consider perception as a higher level process. Perception is a sequence of recognitions where context may partially make up for information not complete enough to permit performance of unique individual recognition. However, the use of context allows the accumulation of an error in identifying a whole sequence in an almost unpredictable manner. This implies that two inputs identified by a perceptual process which appear very much alike may result in vastly different decisions.

It appears that humans identifying isolated sounds are essentially performing simple recognitions. However, when they recognize words and phrases, it is a perception of a whole word or phrase upon its completion. The resultant word and phrase errors are of a random nature and are normally corrected by the situation. The rationale behind seeking a representation that allows these types of errors is that humans designed language, humans use it, and these are the types of errors humans make.

We would define error detection and correction as the filling in of a sound unit within a sequence of sound units when there was

a high degree of doubt as to the unit's actual identity. This amounts to a conditional correction. The fact that context could be used at all times implies that a large number of minor error corrections continuously occur. But, at the degree of uncertainty where a conditional correction is made, it would be possible to consider the next two or three context units for a plausible overall identity. The conditional correction could be noted as a possible major error in the output.

The nonsense or start-over response is a consequence of allowing the error corrections we previously discussed. If we consider the sequence of sound units as specifying individual words, then it would be possible for a correction to take place, continue with decisions leading along an acceptable sequence of sound units until a point is reached where no plausible error correction could be made. At this point, the machine would be thoroughly confused, and to process any further inputs would be nonsense.

The methods described in the next section result from a careful consideration of the previous five characteristics and the concepts expressed in the linguistic strata model. The actual list of comparable and conflicting design constraints induced by our knowledge of natural language and these five constraints is lengthy. The overall conclusion of this evaluation is that the basic information-carrying segment of the acoustic signal must be defined in a context-free sense at several levels. Context-free means the avoidance of problems such as measuring A, but A is followed by B, therefore A could be C if preceded by D, but then it could be . . .

4.0 DATA AND ANALYSIS PROCEDURES

In accordance with our definition of linguistically encoded information as being the same in identical words said by different speakers, five speakers were chosen for their noticeably different manners of speaking. These give speakers each uttered the same ten words in random order (from flashed cards) and were recorded on a high-quality audio tape unit.* Ten multi-syllable words were chosen to contain examples of the general phonetic categories of sounds. Three male speakers were used who may be grossly described as: 1) small man with a slightly nasal sound; 2) medium-sized man with a typical voice, but slower than average enunciation; 3) a large man with a typically full male voice. Two female speakers were used. One woman was small with a strong, concise voice, while the other woman was of medium build with a low, quiet, and almost breathy voice. All five speakers spoke English as their native language.

Three methods were considered for generating a graphical display of the data that would be suitable for the initial visual examination. From time series plots (computer aided), it was not difficult to identify the times that the character of the signal underwent both abrupt and subtle changes. But, due to the complex nature of the speech waveform, it is extremely difficult to discuss the similarities of the subtle waveform changes. Thus, it was decided that a display of the power spectrum as it changed throughout the utterance would be much more useful.

The 50 utterances were converted to sonagrams (Type B) on a KAY Sonagraph, located at Stanford Research Institute, Menlo Park, California.

*The ten words are: mumble, thimble, downward, forward, before, whistled, rudder, shovel, engine, gosling.

Though there are numerous approximations in the hardware realization of this machine, the sonagrams do provide a graphical display which clearly exhibits the overall changes in the spectrum of the signal during the utterance. A more detailed picture of these spectrum changes was obtained with the use of a fast Fourier transform computed over a 30-millisecond interval, with the intervals stepped (overlapping) every six milliseconds throughout each utterance. With the original ten kilocycle-per-second sampling rate, this provided coefficients of 150 components, 33 cycles per second apart, from 33 to 4950 cycles per second. The computed spectra of each utterance (approximately 100 frames of 150 components) were very cumbersome to work with.

The sonagrams rather than the computed spectra were selected for the final analysis. Even though the accuracy with which they can be read is poor, it is much easier to make visual interpretations and comparisons of hypothesized invariant relationships. After an initial subjective inspection of the data, the following methods were chosen in an attempt to insure the objectivity of the complete analysis.

We chose a method of segmentation designed to detect any apparently significant changes in the character of the speech signal. A change in character may be abrupt as the cessation of the voicing of a vowel preceding the pure noise of a fricative or as subtle as the downward trend in the second-format trajectory of a vowel glide. The basis for detecting these is the comparison of the spectrum at t_k with a prediction of the spectrum generated from the two immediately preceding (times t_{k-1}, t_{k-2}) spectra. If there is a discrepancy between the current input and the most recent estimate, and this difference continues, then a segment boundary is

marked and a new prediction is generated as soon as sufficient data becomes available. Abrupt changes are easily detected. The sensitivity to subtle changes is determined by the length of the prediction interval and its relation to the degree of discrepancy required. The process has been carried out by digitizing the sonagrams in data vectors, creating a prediction vector, and finally storing the history of the signal between segment boundaries in terms of state vectors.

The sonagrams represent the spectral energy over the range of 85 cycles per second to 4000 cycles per second. The utterances were from 500 milliseconds to 900 milliseconds in duration (with an approximately 200 millisecond long, 1 kilocycle per second, calibration tone, placed 441 milliseconds before each word). The time scales are 200 milliseconds of real time per inch horizontally and 979 cycles per second per inch vertically. Thought of as a surface, the darkest shadings represent the highest peaks while the absence of shading is a valley of undeterminable depth. The sonagrams were digitized manually by overlaying them with a clear plastic grid having vertical rulings every one-twentieth of an inch and horizontal rulings spaced one-eighth of an inch apart. Each rectangle (1/20th" by 1/8th") in each vertical strip was assigned a number (0, 1, 2, 3) to represent the amount of shading within it (Figure 5). This produces a 32-component vector that represents the distribution of energy (as seen with the sonagram machine) over the speech spectrum in a 10-millisecond period. These will henceforth be referred to as the data vectors.

A prediction vector (the estimate of the next data vector) is formed as a linear projection of the two immediately preceding data vectors. This

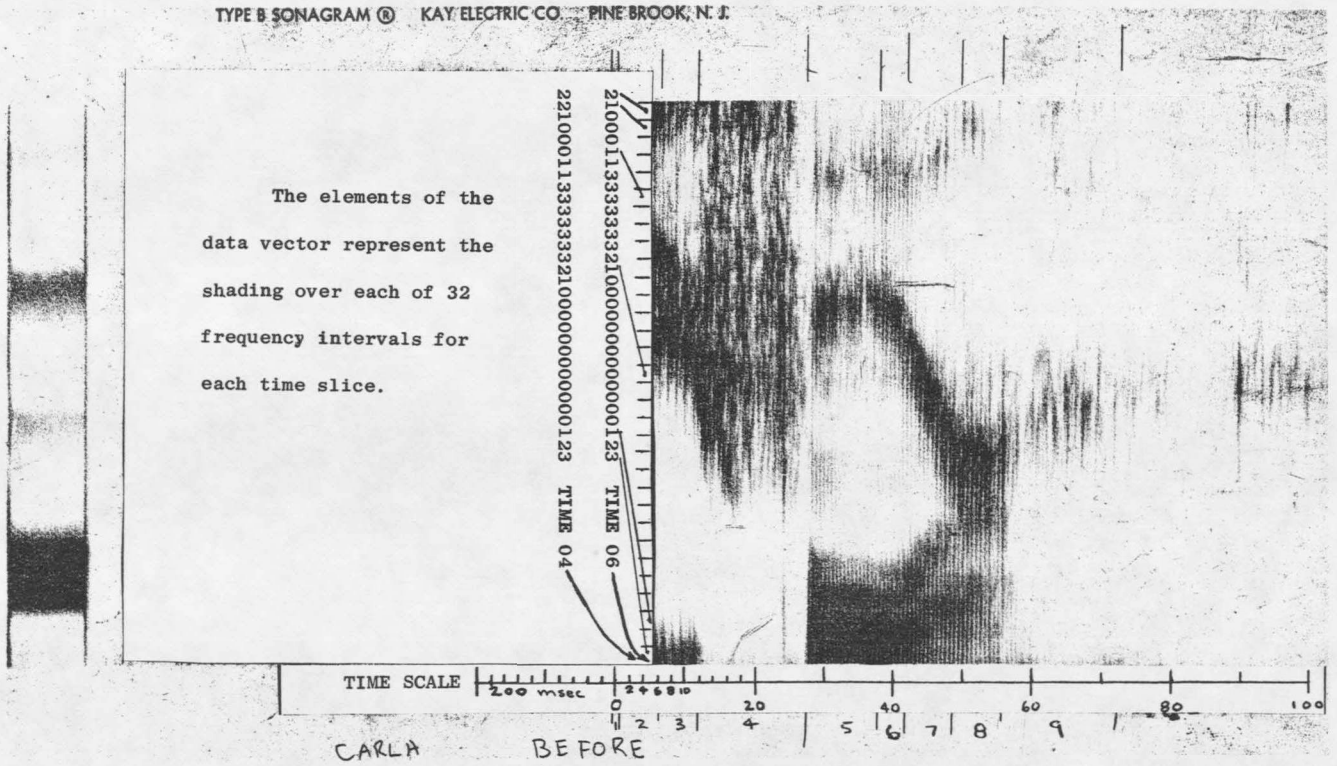


Figure 5. Conversion of Sonagram to Data Vector.

prediction is achieved through the use of shape functions to represent the significant peaks of each data vector (Figure 6). Only the larger three peaks of energy are represented unless there is over 10% of the energy left unrepresented, in which case a fourth peak is allowed. As illustrated in Figure 6, the two preceding data vectors are inspected and linear projections are made to form the prediction vector, which is then reconstructed as a data vector. This is differenced with the current data vector to produce the difference vectors shown in Figure 7. Initial data analysis established the following thresholds for the difference vector that indicate the requirement of a segment boundary.

Thresholds on Difference Vector

1. Silence to anything/anything to silence.

   The difference vector is either a positive or negative image of the last data vector.

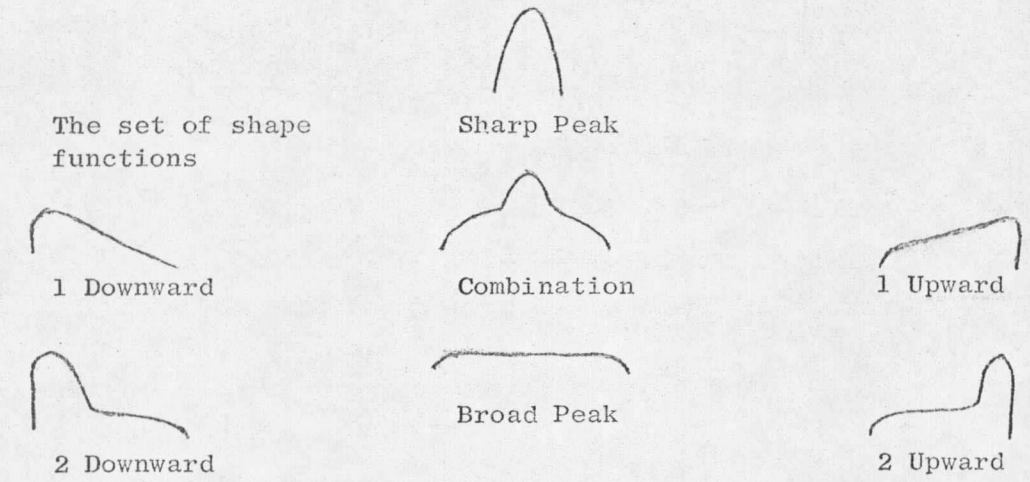
2. All other cases.

   The magnitude of the difference vector larger than five in two consecutive frames and the second frame difference vector elements form one or more significant peaks (with respect to the total sum of the elements).

Two additional important segmentation indicators are determined directly from the data vectors. The first is any change in the number of significant peaks, or the merging or splitting of peaks in the incoming data vectors. These are necessary conditions for creating comparable segments across speakers. The second is a detection of long pure-noise segments.

The data vectors are represented with the aid of shape functions in

order to generate the prediction vector.



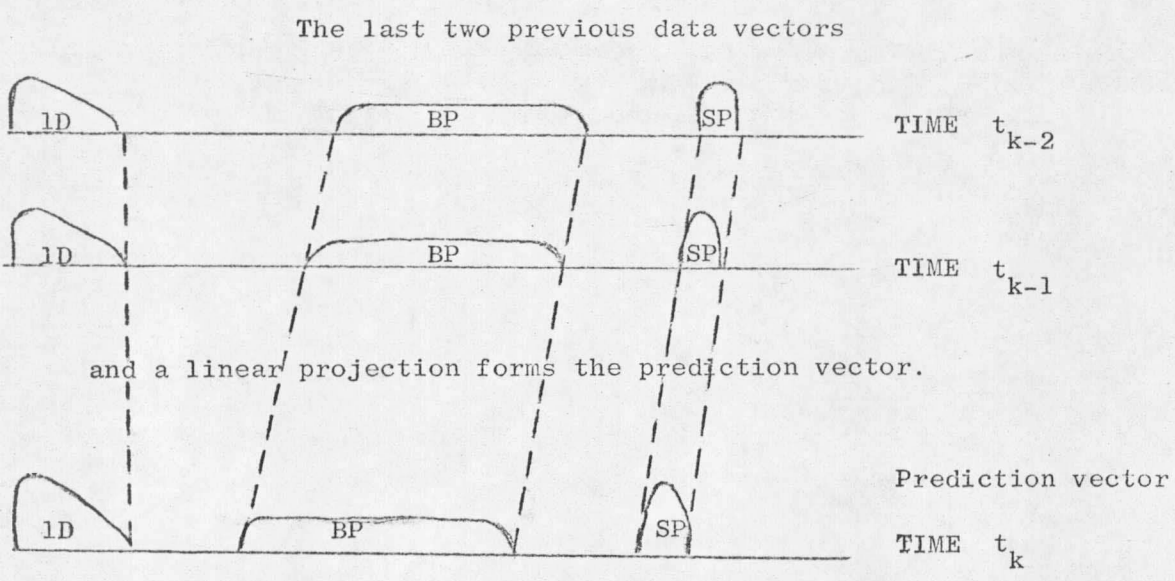


Figure 6. Generation of Prediction Vector

Segmentation within /before/ spoken by CH

TIME	DATA VECTOR	
36	333321000000000000023210000110000	PREDICTION VECTOR
		DIFFERENCE VECTOR
37	333321000000000000023210000110000	---

38	333321000000000000023210000110000	333321000000000000023210000110000
		000000000000000000000000000000000
		SEGMENT
39	333321000000000001232100001100000	333321000000000000023210000110000
		000000000000000001211110001010000
40	333321000000000012321000000110000	333321000000000000023210000110000
		000000000000000012302210000000000
41	333321000000000123210000000110000	333321000000000123210000000011000
		000000000000000000000000000101000
42	333321000000001232100000000011000	333321000000001232100000000110C0
		000000000000000000000000000101000
		SEGMENT
43	333332100000012321000000000000000	333321000000012321000000000001100
		000011100000000000000000000001100
44	333333210000123210000000000100000	333321000000123210000000000000110
		000012210000000000000000000100110
45	333333321000123210000000000000000	333333321001232100000000000000000
		000000000000000000000000000000000
46	333333332012321000000000000000000	333333332112321000000000000000000
		000000001000000000000000000000000
47	233332221123210000000000000000000	333333333223210000000000000000000
		100001112100000000000000000000000
48	2333322222332100000000000000000000	Peaks merged - SEGMENT

Figure 7. Segmentation with Data, Prediction, and Difference Vectors.

This can be noted when three or more data vectors contain a large number of adjacent non-zero components that vary from frame to frame. These segments will be the long noise-like unvoiced sounds. Figure 8 illustrates the seqquence of segmentations produced in the word /before/.

In order to preserve the history of the speech signal during each of the time segments, a state vector is created. Since by definition no significant changes occurred in the signal's character between two segment boundaries, we need only linearly interpolate between the initial data vector of the segment and the final data vector. The elements of the state vector are illustrated in Figure 9. Figure 10 lists a state vector's record for the utterance displayed in Figure 8. Various combinations of calculations may be selected as the elements to be recorded in the state vector. Those used here correspond to the initial data vector and the changes that accur in it in reaching the final data vector. With this choice of elements, a glance at the state vector indicates immediately the dynamic changes occurring during the state. There are also certain cases where the state vector elements are quite redundant. For example, in a one-peak segment the initial energy and final energy are recorded both for the total segment and for the individual peak. There being only one peak, the total energy and the single peak energy measurements are identical. The state vector may be generated immediately after the specification of the terminal boundary of the segment is made. It then becomes the physical record of the sound unit for all further computations.

Figure 11 displays the acoustic cues in the word /before/ from a composite of our five samples. The composite is an average of the characteristics

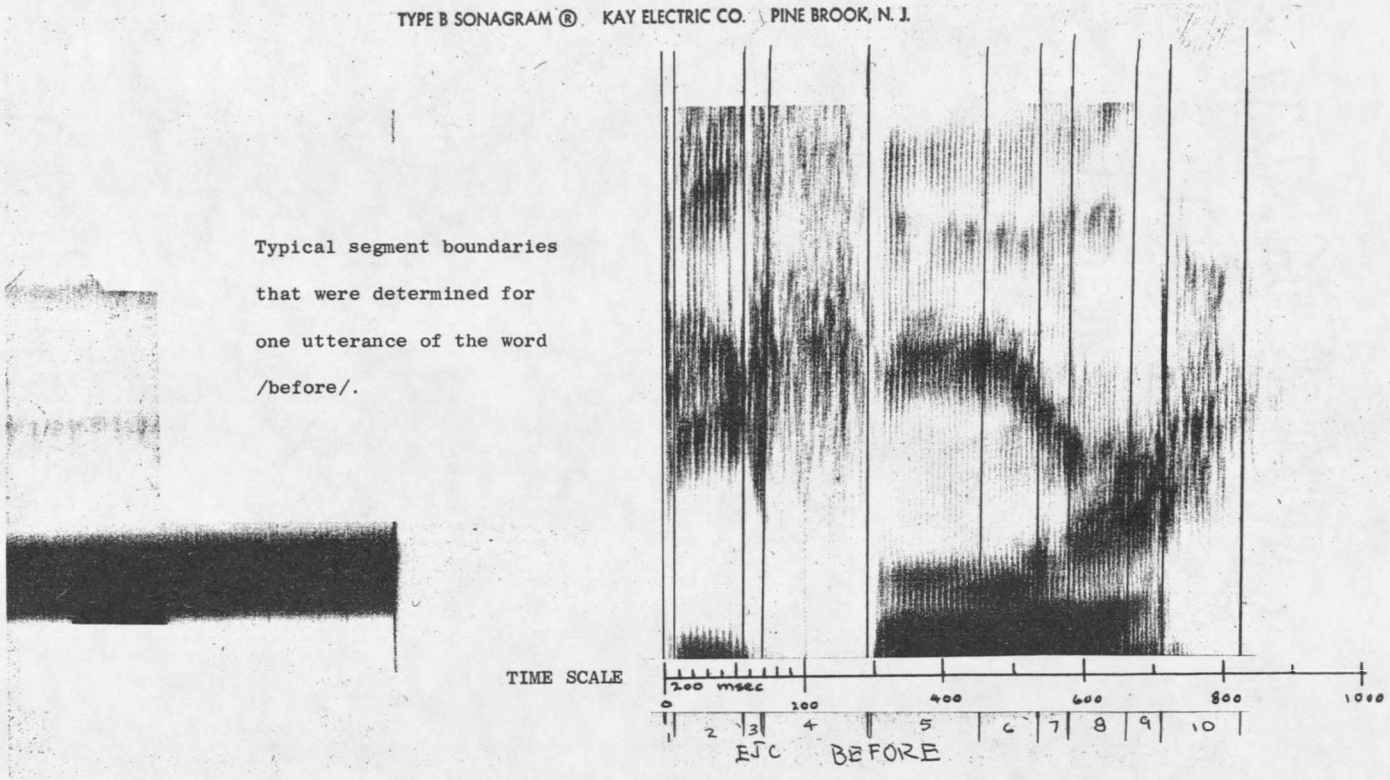


Figure 8. Pictorial Example of Segmentation of /Before/.

The State Vector Components

Parameters of Total Segment					Parameters of Individual Peaks							
noise /flag	# of /peaks	initial total /energy	final total /energy	time /length	peak /total	peak /change	begin /shape	end /shape	lower /boundary	lower boundary /change	upper /boundary	upper boundary /change

// Since this portion contains the specific information about // each peak, it is repeated for each significant peak.

Typical boundary data vectors		The corresponding state vectors	
TIME		STATE	STATE
00 01	000000000012222222222111100000000 000000000012222222222111100000000	1	1 0/1/25/25/01///25/ 0/BP/BP//11/ 0//25/ 0//
02 08	320000000012223333332222211110000 320000000000123333333222222211000	2	2 0/2/42/43/06/// 5/ 0/1D/1D// 1/ 0// 2/ 0// ///37/ 1/ C/1D//11/ 2//28/ 1//
09 14	320000000000123333333222222211000 320000000001233333322221111111100	3	3 0/2/43/43/05/// 5/ 0/1D/1D// 1/ 0// 2/ 0// ///38/-2/1D/1D//13/-1//29/ 1//
15 29	000000000000122222222222222111111N 000000000000122222222222222111111N	4	4 1/1/34/34/14///34/ 0/BP/BP//13/ 0//32/ 0//

Figure 9. The Formation of the State Vectors

STATE VECTOR RECORD

Utterance CH-5

No.1 0 / 1 / 43 / 43 / 01 /// 43 / 0 / BP / BP // 11 / 0 // 32 / 0 //
// / / / / // / // / //
// / / / / // / // / //

No.2 0 / 3 / 39 / 35 / 04 /// 6 / 0 / 1D / 1D // 1 / 0 // 3 / 0 //
// 26 / 0 / BP / BP // 15 / 2 // 27 / 0 //
// 7 / -4 / 1U / 1U // 29 / 2 // 32 / 0 //

No.3 0 / 3 / 35 / 36 / 05 /// 6 / 0 / 1D / 1D // 1 / 0 // 3 / 0 //
// 26 / 1 / BP / BP // 17 / -4 // 27 / -4 //
// 3 / 0 / 1U / 1U // 31 / 0 // 32 / 0 //

No.4 1 / 1 / 48 / 48 / 14 /// 48 / 0 / BP / BP // 11 / 0 // 32 / 0 //
// / / / / // / // / //
// / / / / // / // / //

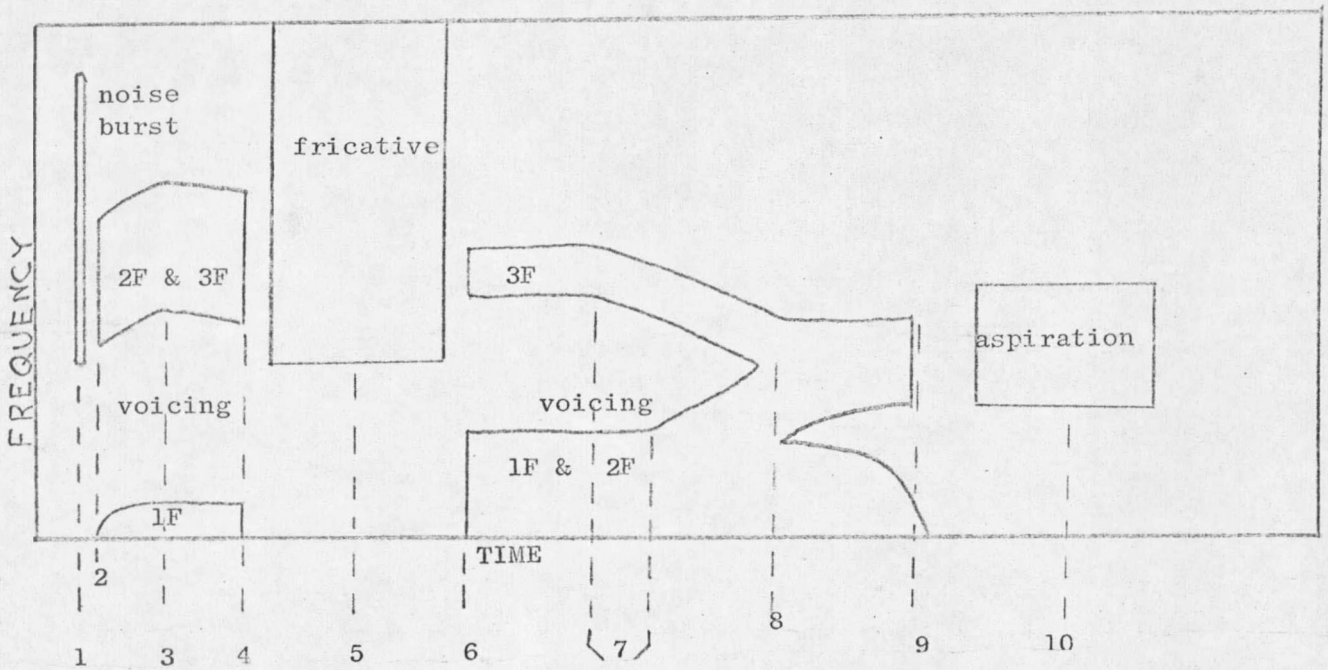
No.5 0 / 3 / 24 / 24 / 10 /// 15 / 0 / 1D / 1D // 1 / 0 // 6 / 0 //
// 7 / 0 / SP / SP // 19 / 0 // 22 / 0 //
// 2 / 0 / BP / BP // 28 / 0 // 29 / 0 //

No.6 0 / 3 / 24 / 26 / 03 /// 15 / 0 / 1D / 1D // 1 / 0 // 6 / 0 //
// 7 / 2 / SP / SP // 19 / -4 // 22 / -3 //
// 2 / 0 / BP / BP // 28 / 1 // 29 / 1 //

No.7 0 / 2 / 27 / 30 / 04 /// 18 / 3 / 1D / 2D // 1 / 0 // 7 / 2 //
// 9 / 0 / SP / SP // 14 / -4 // 18 / -4 //
// / / / / // / // / //

Figure 10. The First Seven State Vectors for Speaker CH Uttering /Before/.

Composite of the acoustic cues in the word /before/



1 The initial explosive noise burst of the /b/. Only one speaker was observed to voice the /b/ in this initial position.

2, 3 The vowel has a strong onset of 1F that continues until 4. The 2F and 3F are very close with an abrupt path change at 3.

4, 5 The abrupt cessation of voicing followed by the pure noise of the fricative. The lower boundary of the noise appears below the lower 2F boundary.

6 The termination of the long fricative is signaled by the strong onset of the voiced 1F and 2F.

7 The 3F travels downward and the 2F comes upwards towards the final merge at 8. The exact time that the 3F and 2F make definite moves varies considerably across speakers.

8 A final steady state portion that concludes the voicing of the utterance.

10 This is the release of any remaining breath at the termination of the utterance and consequently the range and intensity are relatively unimportant.

Figure 11. Illustration of Some Common Acoustic Cues.

contained in the five utterances. The segments have been artificially separated to accentuate the boundary difference between segments. Figure 12 compares one speaker's realization of /before/ to the composite of the five speakers' realizations. In a very real sense, the relationship between the single realization and the composite is that of phonetic to phonemic. In comparing these two sequences it is apparent that many of the distinctive elements are the transitions between the state vectors. In order to produce an equivalence between these two sequences of states, Production Modes and Relative Oppositions were created.

The Production Modes and Relative Oppositions are a result of the need to represent utterances as sequences of elements and the relationships between these elements. The Production Mode (PM) is one of eleven classes into which the state vectors may be divided. The eleven PM's are listed in Figure 13. The PM classes are gross enough in nature to overcome many characteristics of speaker variation. For instance, the vowels with three significant formants all fall within the same PM. Another PM contains all voiceless fricatives greater than 40 milliseconds in duration. The Relative Opposition (RO) is a set of calculations that provide a detailed record of the distinctive changes that characterize the transitions between neighboring PM's. The RO contains more information than is available at the instant of segment boundary determination. This results from the fact that RO computations are based both on boundary phenomena and on the total signal history throughout both states. The calculation of RO_i (i being the time index) requires the determination of PM_i and PM_{i+1}; the resultant RO_i is associated with the transition from PM_i to PM_{i+1}. The 42 detailed calculations that

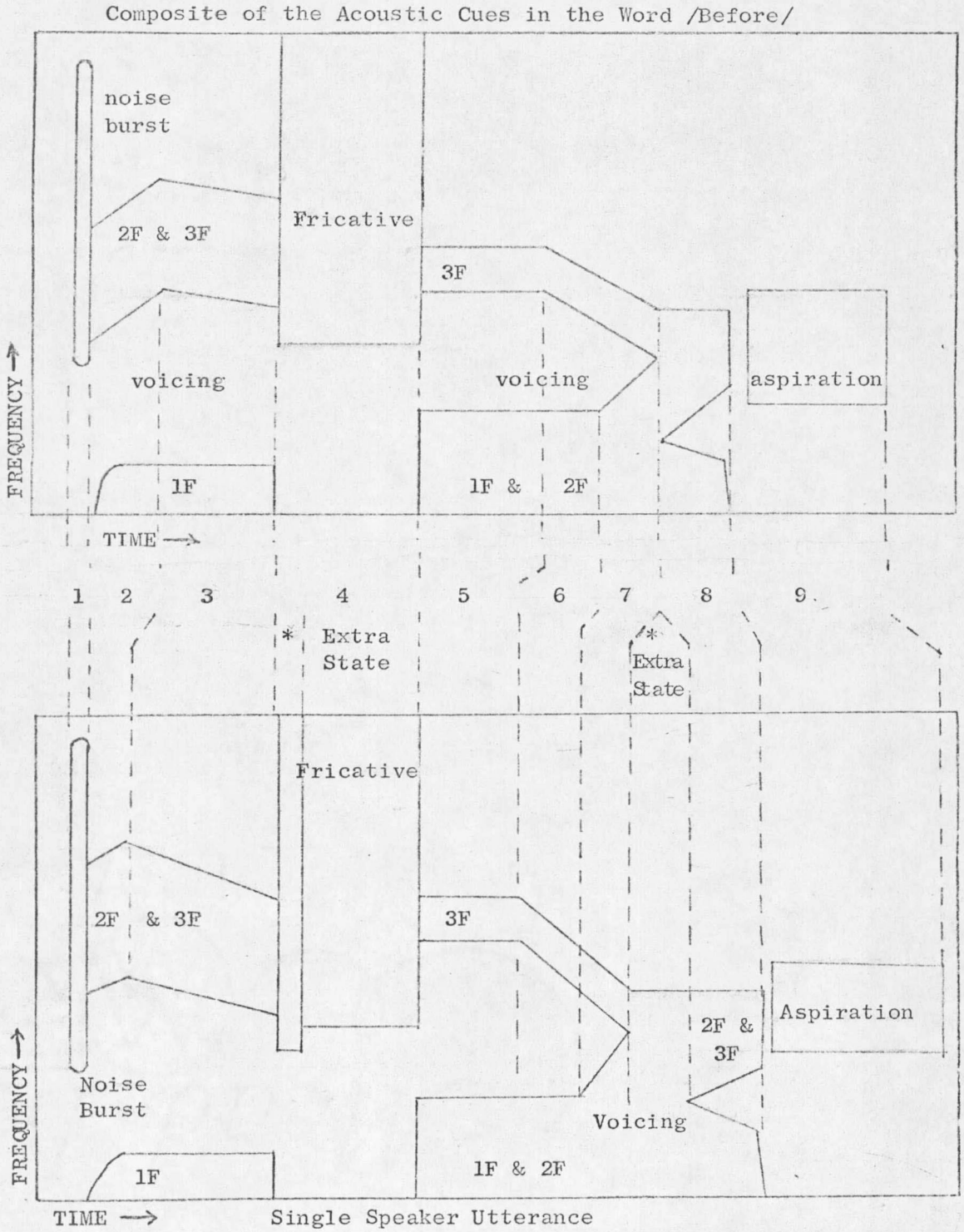


Figure 12. Comparison of Single Speaker Utterance to Composite of /Before/.

Production Modes

	Duration	No. of Peaks	Power Distribution	Comments
PM 1	> 40 msec	1	NA	Voiced Continuents
PM 2	> 40 msec	2	NA	Voiced Continuents
PM 3	> 40 msec	3	NA	Voiced Continuents
PM 4	> 40 msec	1	NA	Noise Flag
PM 5	> 40 msec	2	< 10% / > 90%	Voiced Construents
PM 6	< 40 msec	1	NA	Mostly Stops
PM 7	< 40 msec	2	< 10% / > 90%	Voiced Noise Burst
PM 8	< 40 msec	2	> 10% / < 90%	Transient Segments
PM 9	< 40 msec	3	NA	Transient Segments
PM 10 (ϕ)	< 100 msec	quiet		
PM 11 (S)	> 100 msec	silence		

N.B. There is a certain symmetry that is obtained by listing PM's that differ only in duration.

PM 4 ~ PM 1 with Noise Flag

PM 6 ~ Short PM 1

PM 7 ~ Short PM 5

PM 8 ~ Short PM 2

PM 9 ~ Short PM 3

PM 5 ~ PM 2 with Special Energy Distribution

Figure 13. Classes of State Vectors in the Production Modes

compose the total RO_i are denoted as the set $\{ro_j\}_i$, where i is the time index, and j=1, 42 indexes the individual results recorded on the RO_i data sheet of Fig. 14. The extensive algorithms for the $\{ro_j\}$ calculations are exhibited in Fig. 15 as a sequence of illustrations that detail all cases. The results of all $\{ro_j\}$ calculations are either numerical or indeterminant. The letter "I" indicates an indeterminant calculation that either makes no sense to perform or has no counterpart in the PM transition.

The use of the indeterminant I as the value of an ro_j is an integral part of the RO_i algorithms. The RO_i records continuous and discontinuous features of the signal across the PM boundary. The discontinuous case is that of a formant terminating at a boundary without intersecting another formant; the most common example of this is the cessation of voicing; other examples occur in the second and third formants of medial and terminal nasals. The RO_i that records this discontinuous phenomenon has a subset of $\{ro_j\}$ that appear to represent a peak not actually present. In the case of termination, no comparison can be made between the formant's dynamic character in the PM_i and PM_{i+1} state. Consequently, the information requiring comparisons of the formant in the PH_i and its character in the PM_{i+1} is recorded as indeterminant. The initiation of a formant causes the inverse to be recorded. A further result of this approach is that an RO_i representing the termination of a three-formant vowel contains three elements, while nothing is there. The worst case possible would be the termination of three formants and the initiation of three more, requiring the RO_i to have six major elements (peak sections). We have found that three peak sections of the RO_i were sufficient in this data.

Relative Opposition Data Sheet

PM_i 2 — PM_i SEQ NO. 6

PM_{i+1} 3 — SAMPLE WORD 5-PL

ro_1	Duration Ratio 21/6											
ro_2	Power Ratio 21/23											
ro_{3-6}	Power Distribution 71	29		/	48	26	26					
ro_9	Merge; Split 01	Boundary Revisions										
ro_{10-15}	DESTROYED	L1 0	U1 1	L2 1	U2 1	L3 0	U3 0					
ro_{16-21}	CREATED	L1 0	U1 1	L2 1	U2 1	L3 1	U3 1					

PK-1 Information

ro_{22}	Abrupt Initiation (Power and Shape) 0				
ro_{23}	Abrupt Termination (Power and Shape) 0				
ro_{24-25}	Abrupt Boundary Discontinuity	LB Step 0		UB Step I	
ro_{26-28}	Shapes I/1D	Slopes 0/0		Slopes I/-.16	

PK-2 Information

ro_{29}	Abrupt Initiation (Power and Shape) 0				
ro_{30}	Abrupt Termination (Power and Shape) 0				
ro_{31-32}	Abrupt Boundary Discontinuity	LB Step I		UB Step 0	
ro_{33-35}	Shapes I/BP	Slopes I/+.16		Slopes 0/+.16	

PK-3 Information

ro_{36}	Abrupt Initiation (Power and Shape) 0				
ro_{37}	Abrupt Termination (Power and Shape) 0				
ro_{38-39}	Abrupt Boundary Discontinuity	LB Step 0		UB Step 0	
ro_{40-42}	Shapes BP/BP	Slopes -.1/-.3		Slopes -.1/-.3	

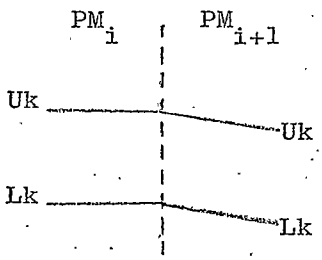
Figure 14. The RO_i Calculation Recording Form

The algorithms for the computation of the $\{ro_j\}$ are most easily presented as a number of specific cases. All the more general cases are straightforward combinations of these specific cases. The following notation has been adapted for the sake of simplicity.

1. Lower and upper boundaries of peaks are designated as L1, L2, L3 and U1, U2, U3; except in the basic single-peak cases where the Kth peak boundaries are designated as Lk and Uk.
2. The $\{ro_j\}$ elements dealing with duration, total power and power distribution are not included, since there are no ambiguities in the calculation.
3. The shape of each peak is designated as SHk_i and SHk_{i+1} in terms of the shape functions used for the state vectors (Figure 6).
4. The slope of each boundary is designated as SLk_i and SLk_{i+1} for lower boundaries and SUk_i and SUk_{i+1} for upper boundaries.
5. The discontinuous boundary steps that have a numerical value are designated $J_{\text{boundary PM}_i\ /\ \text{boundary PM}_{i+1}}$.

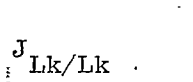
Figure 15. Algorithms for the Generation of the RO_i.

Basic Continuous



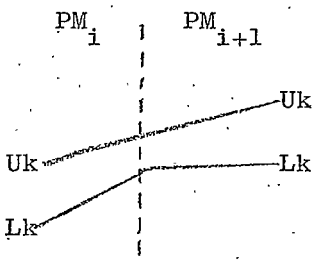
PK-k

SHk_i / SHk_{i+1}



$J_{Lk/Lk}$

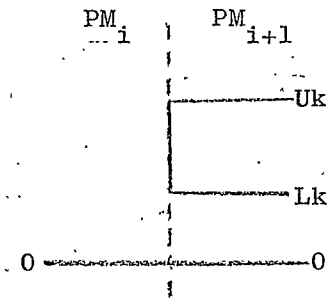
SLk_i / SLk_{i+1}



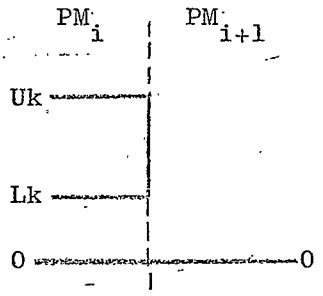
$J_{Uk/Uk}$

SUk_i / SUk_{i+1}

Basic Discontinuous



PK-k Abrupt Initiation

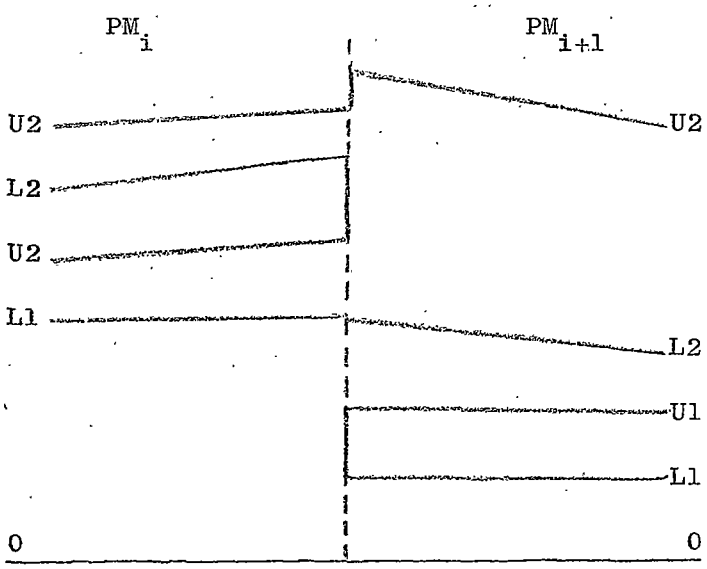


PK-k Abrupt Termination

$\frac{I}{I/SHk_{i+1}}$ $\frac{I}{I/SLk_{i+1}}$ $\frac{I}{I/SUk_{i+1}}$ $\frac{SHk_i/1}{}$ $\frac{I}{SLk_i/I}$ $\frac{I}{SUk_i/I}$

Figure 15 continued.

Combination of Continuous and Discontinuous



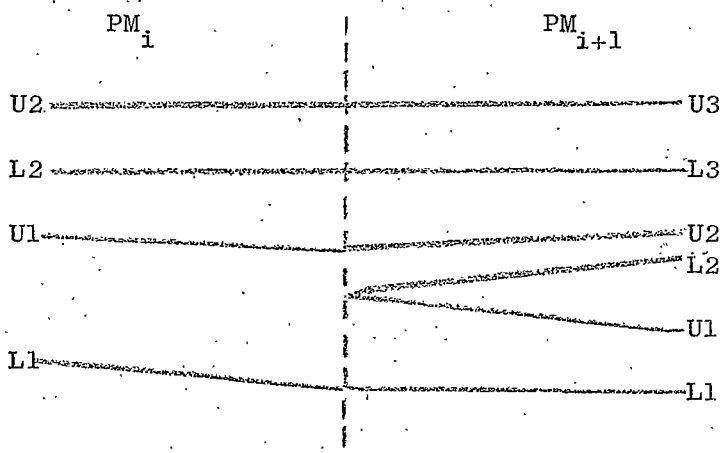
Significant Elements

Abrupt Initiation

ro_{24-25}	PK-1	I	I
ro_{26-28}	I/SHI_{i+1}	$I/SL1_{i+1}$	$I/SU1_{i+1}$
ro_{31-32}	PK-2	$J_{L1/L2}$	$J_{U2/U2}$
ro_{33-35}	$I/SH2_{i+1}$	$SL1_i/SL2_{i+1}$	$SU2_i/SU2_{i+1}$

Figure 15 continued.

Merge/Split 01



Significant Elements of RO_i

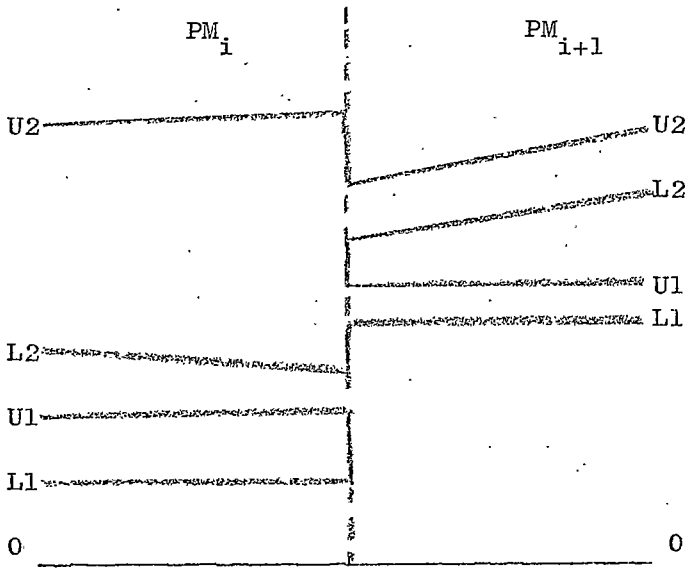
ro_9 m/s 01

		L1	U1	L2	U2	L3	U3
ro_{10-15}	D	0	1	1	1	0	0
ro_{16-21}	C	0	1	1	1	1	1

ro_{24-25}	PK-1	$J_{L1/L1}$	I
ro_{26-28}	$I/SH1_{i+1}$	$SL1_i/SL1_{i+1}$	$I/SU1_{i+1}$
ro_{31-32}	PK-2	I	$J_{U1/U2}$
ro_{33-35}	$I/SH2_{i+1}$	$I/SL2_{i+1}$	$SU1_i/SU2_{i+1}$
ro_{38-39}	PK-3	$J_{L2/L3}$	$J_{U2/U3}$
ro_{40-42}	$SH2_i/SH3_{i+1}$	$SL2_i/SL3_{i+1}$	$SU2_i/SU3_{i+1}$

Figure 15 continued.

Combination of Continuous and Discontinuous



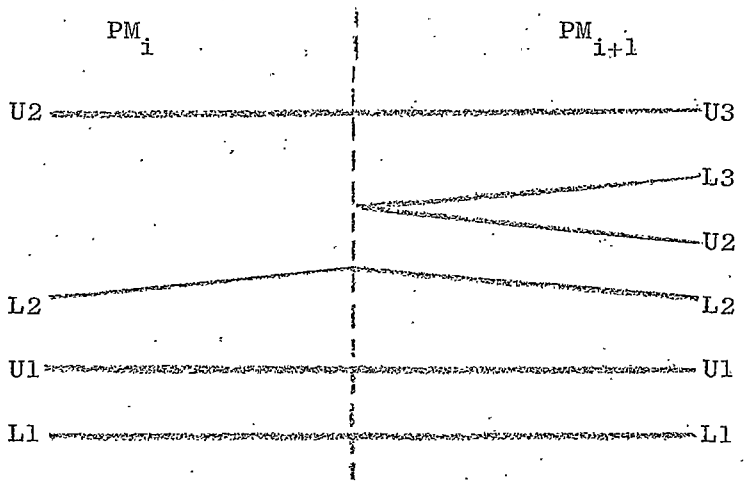
Significant Elements of RO_i

Abrupt Termination

ro_{24-25}	PK-1	I	I
ro_{26-28}	$SH1_i/I$	$SL1_i/I$	$SU1_i/I$
ro_{31-32}	PK-2	$J_{L2/L1}$	I
ro_{33-35}	$I/SH1_{i+1}$	$SL2_i/SL1_{i+1}$	$I/SU1_{i+1}$
ro_{38-39}	PK-3	I	$J_{U2/U2}$
ro_{40-42}	$I/SH2_{i+1}$	$I/SL2_{i+1}$	$SU2_i/SU2_{i+1}$

Figure 15 continued.

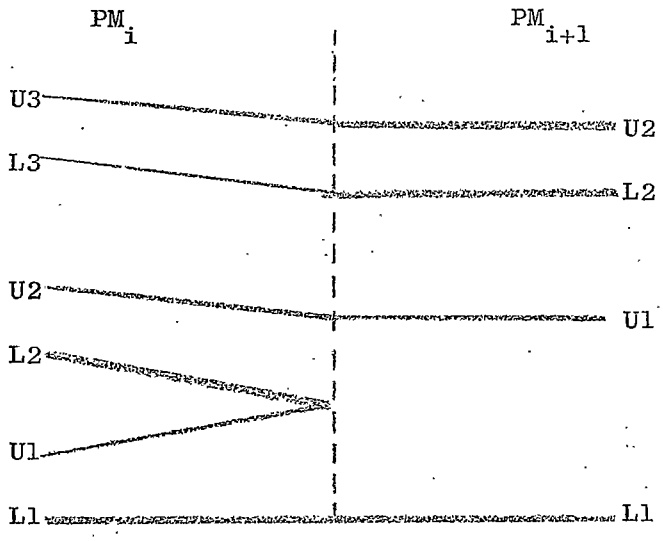
Merge/Split 01



Significant Elements of RO_i

ro_9	m/s	01					
		L1	U1	L2	U2	L3	U3
ro_{10-15}	D	0	0	0	1	0	0
ro_{16-21}	C	0	0	0	1	1	1
ro_{24-25}		PK-1		$J_{L1/L1}$		$J_{U1/U1}$	
ro_{26-28}		$SH1_i/SH1_{i+1}$		$SL1_i/SL1_{i+1}$		$SU1_i/SU1_{i+1}$	
ro_{31-32}		PK-2		$J_{L2/L2}$		I	
ro_{33-35}		$I/SH2_{i+1}$		$SL2_i/SL2_{i+1}$		$I/SU2_{i+1}$	
ro_{38-39}		PK-3		I		$J_{U2/U3}$	
ro_{40-42}		$I/SH3_{i+1}$		$I/SL3_{i+1}$		$SU2_i/SU3_{i+1}$	

Merge/Split 10

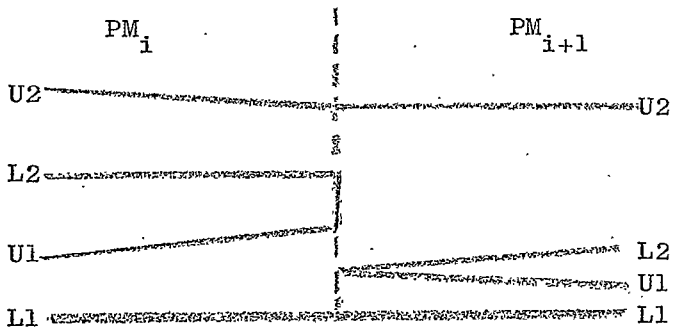


Significant Elements of RO$_i$

ro$_9$	m/s	10					
		L1	U1	L2	U2	L3	U3
ro$_{10-15}$	D	0	1	1	1	1	1
ro$_{16-21}$	C	0	1	1	1	0	0
ro$_{24-25}$		PK-1		J$_{L1/L1}$		J$_{U2/U1}$	
ro$_{26-28}$		I/SH1$_{i+1}$		SL1$_i$/SL1$_{i+1}$		SU2$_i$/SU1$_{i+1}$	
ro$_{31-32}$		PK-2		J$_{L3/L2}$		J$_{U3/U2}$	
ro$_{33-35}$		SH3$_i$/SH2$_{i+1}$		SL3$_i$/SL2$_{i+1}$		SU3$_i$/SU3$_{i+1}$	

Figure 15 continued.

Merge/Split 01



Significant Elements of RO$_i$

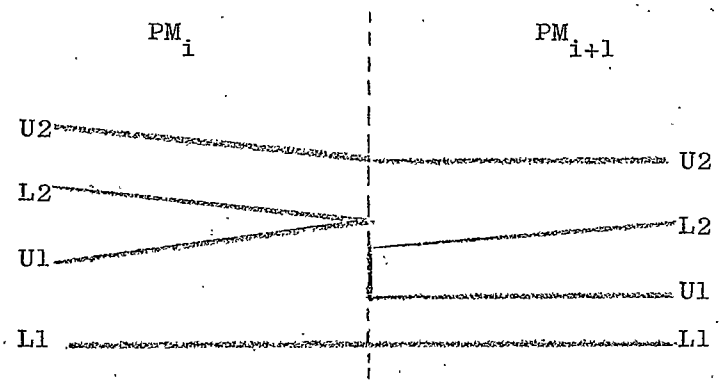
ro$_9$ m/s 01

		L1	U1	L2	U2	L3	U3
ro$_{10-15}$	D	0	1	1	0	0	0
ro$_{16-21}$	C	0	1	1	0	0	0

ro$_{24-25}$	PK-1	J$_{L1/L1}$	I
ro$_{26-28}$	I/SH1$_{i+1}$	SL1$_i$/SL1$_{i+1}$	I/SU1$_{i+1}$
ro$_{31-32}$	PK-2	I	J$_{U2/U2}$
ro$_{33-35}$	I/SH2$_{i+1}$	I/SL2$_{i+1}$	SU2$_i$/SU2$_{i+1}$

Figure 15 continued.

Merge/Split 10



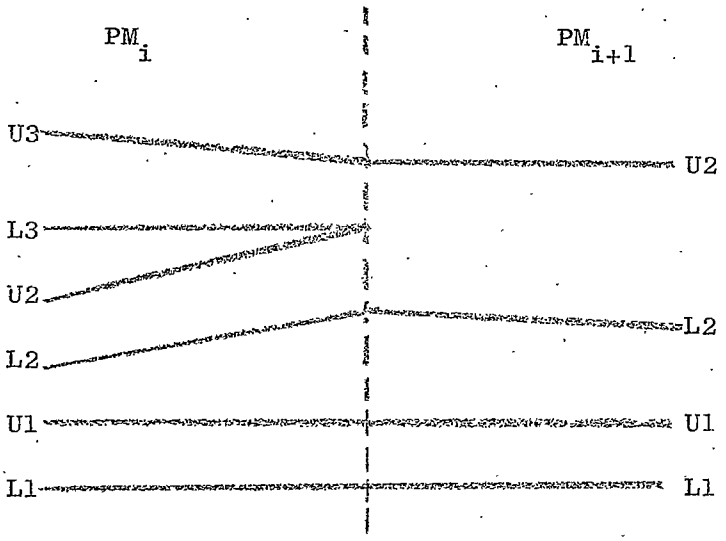
Significant Elements of RO$_i$

ro$_9$	m/s	10					
		L1	U1	L2	U2	L3	U3
ro$_{10-15}$	D	0	1	1	0	0	0
ro$_{16-21}$	C	0	1	1	0	0	0

ro$_{24-25}$	PK-1	$J_{L1/L1}$	I
ro$_{26-28}$	I/SH1$_{i+1}$	SL1$_i$/SL1$_{i+1}$	I/SU2$_{i+1}$
ro$_{31-32}$	PK-2	I	$J_{U2/U2}$
ro$_{33-35}$	I/SH2$_{i+1}$	I/SL2$_{i+1}$	SU2$_i$/SU2$_{i+1}$

Figure 15 conintued.

Merge/Split 10

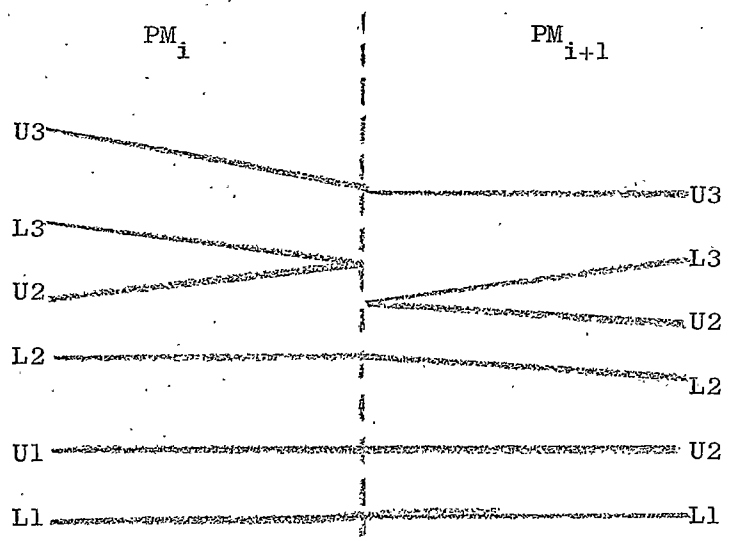


Significant Elements of RO_i

ro_9	m/s	10					
		L1	U1	L2	U2	L3	U3
ro_{10-15}	D	0	0	0	1	1	1
ro_{16-21}	C	0	0	0	1	0	0
ro_{24-25}		PK-1		$J_{L1/L1}$		$J_{U1/U1}$	
ro_{26-28}		$SH1_i/SH1_{i+1}$		$SL1_i/SL1_{i+1}$		$SU1_i/SU1_{i+1}$	
ro_{31-32}		PK-2		$J_{L2/L2}$		$J_{U3/U2}$	
ro_{33-35}		$I/SH2_{i+1}$		$SL2_i/SL2_{i+1}$		$SU3_i/SU2_{i+1}$	

Figure 15 continued.

Merge/Split 11



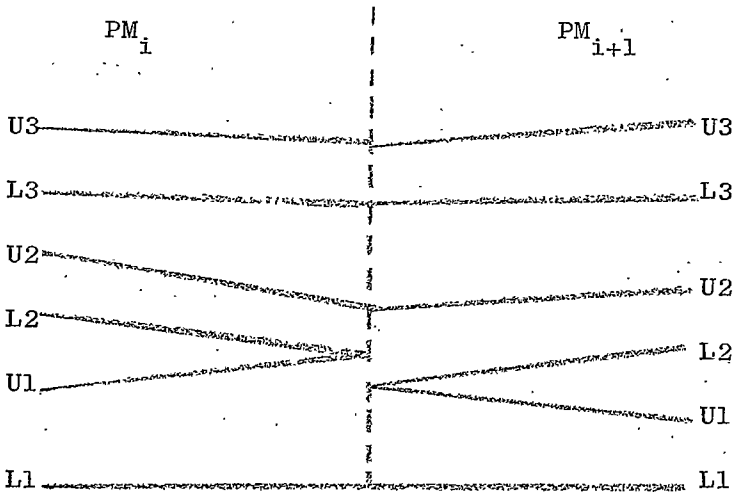
Significant Elements of RO_i

ro_9 m/s 11

		L1	U1	L2	U2	L3	U3
ro_{10-15}	D	0	0	0	1	1	0
ro_{16-21}	C	0	0	0	1	1	0

ro_{24-25}	PK-1	$J_{L1/L1}$	$J_{U1/U1}$
ro_{26-28}	$SH1_i/SH1_{i+1}$	$SL_i/SL1_{i+1}$	$SU1_i/SU1_{i+1}$
ro_{31-32}	PK-2	$J_{L2/L2}$	I
ro_{33-35}	$I/SH2_{i+1}$	$SL2_i/SL2_{i+1}$	$I/SU2_{i+1}$
ro_{38-39}	PK-3	I	$J_{U3/U3}$
ro_{40-42}	$I/SH3_{i+1}$	$I/SL3_{i+1}$	$SU3_i/SU3_{i+1}$

Merge/Split 11



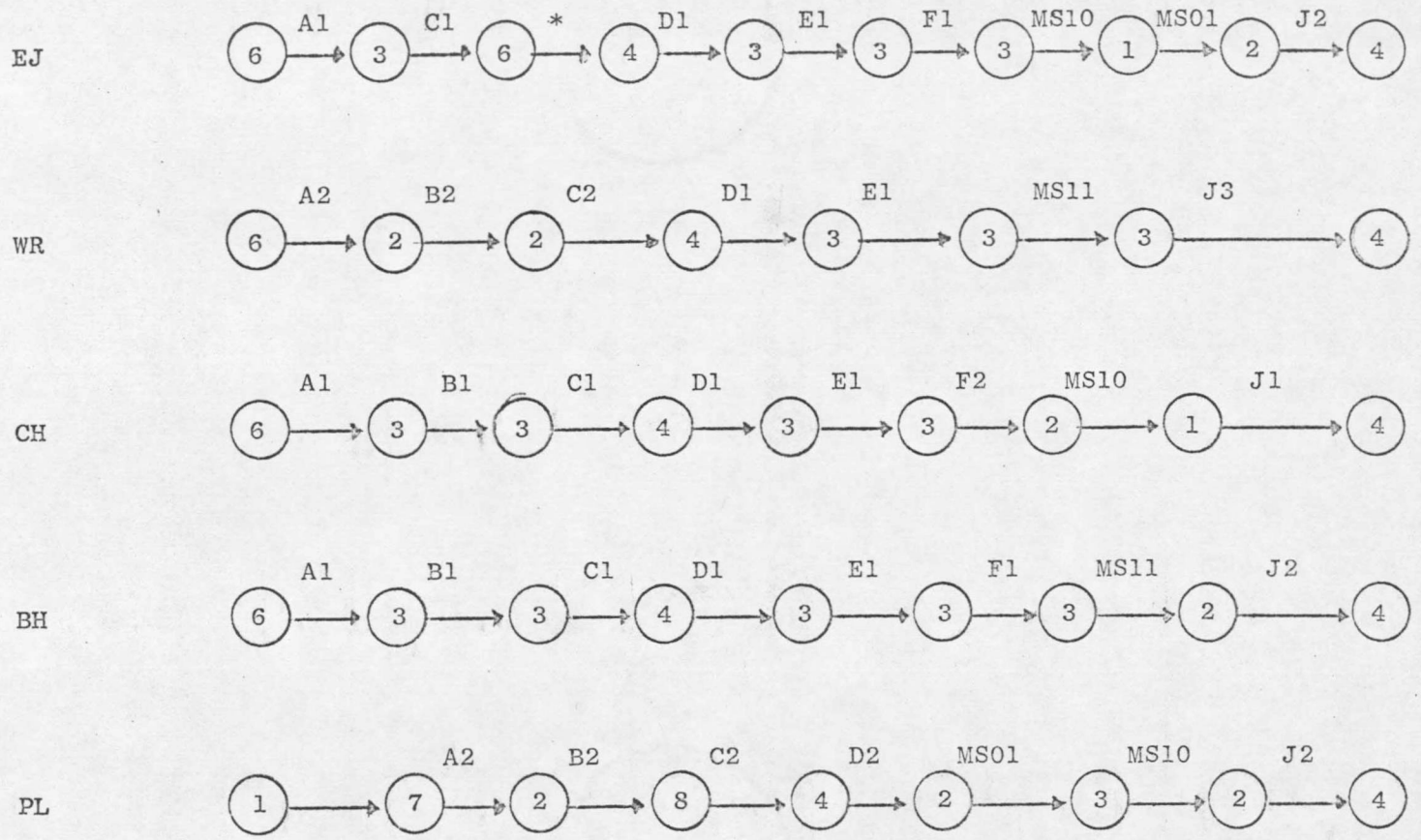
Significant Elements of RO_i

ro_9	m/s	11					
		L1	U1	L2	U2	L3	U3
ro_{10-15}	D	0	1	1	0	0	0
ro_{16-21}	C	0	1	1	0	0	0

ro_{24-25}	PK-1	$J_{L1/L1}$	I
ro_{26-28}	$I/SH1_{i+1}$	$SL1_i/SL1_{i+1}$	$I/SU1_{i+1}$
ro_{31-32}	PK-2	I	$J_{U2/U2}$
ro_{33-35}	$I/SH2_{i+1}$	$I/SL1_{i+1}$	$SU2_i/SU2_{i+1}$
ro_{38-39}	PK-3	$J_{L3/L3}$	$J_{U3/U3}$
ro_{40-42}	$SH3_i/SH3_{i+1}$	$SL3_i/SL3_{i+1}$	$SU3_i/SU3_{i+1}$

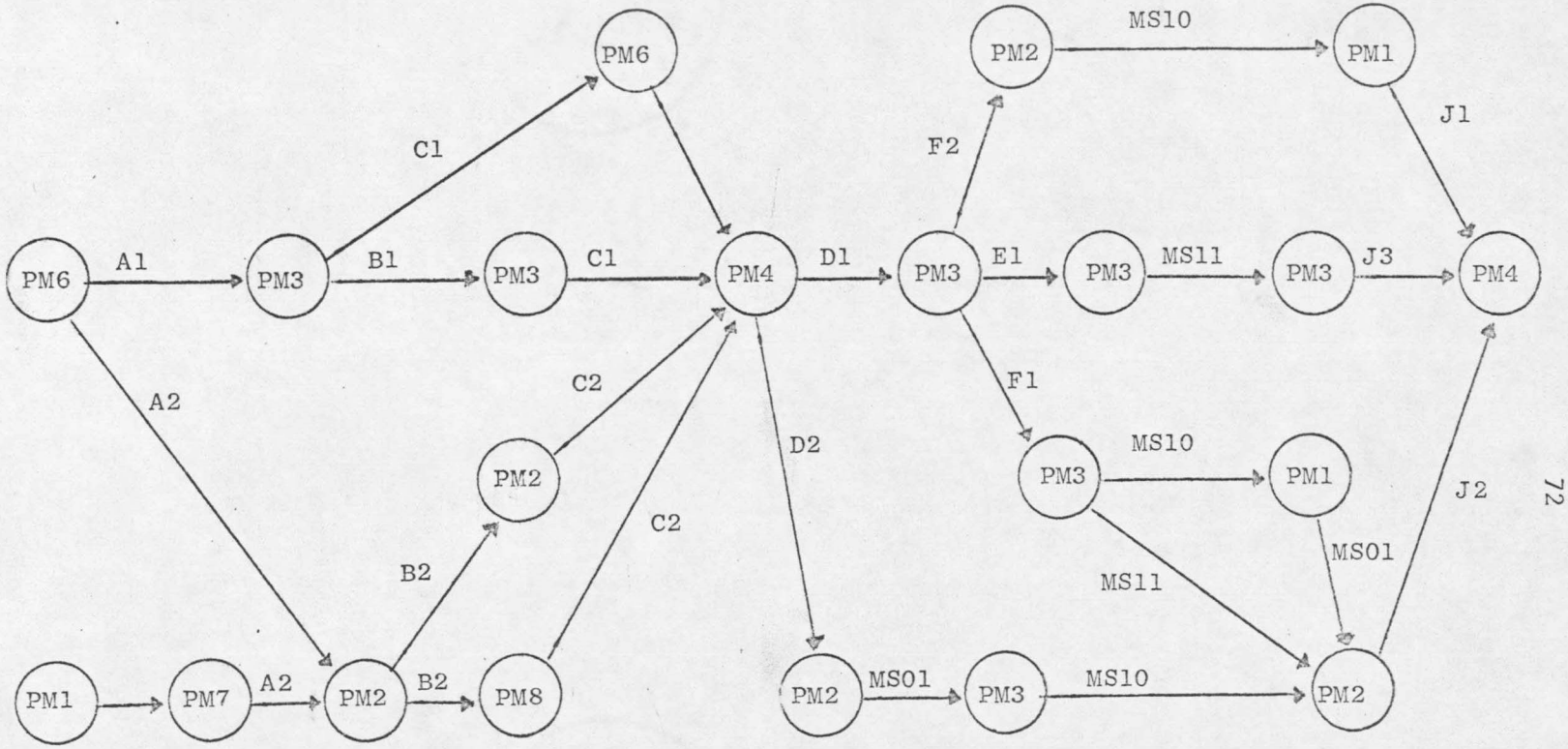
The case of continuous formants presents its own special problems. A single formant that changes in dynamic characteristics is easily recorded. The cases where formants intersect or split is more difficult; it provides more formant boundaries in the following PM. The merge-split portion of the RO_i is essentially only a processing aid. While the subset of $\{ro_j\}$ under the merge-split category does provide a characterization of the exact type of merge-split, its main function is to aid in finding the correct boundaries for comparison of slope changes. There are various numbers of indeterminant ro_j elements in each type of merge-split.

Each utterance of the data set may be plotted as a sequence of RO's and PM's. When this is done for each particular word there appear to be particular PM's that function as tiepoints. The PM-RO sequences generated by the five speakers for the utterance /before/ are exhibited in Figure 16. The tiepoint in this particular word exists in the voiceless fricative /f/ in the medial position and the terminal aspiration. The RO's are shown as links connecting the PM's. The PM's that have similar RO's linking them may be superimposed, as shown in Figure 17. The representation of the terminal /r/ of the utterance /before/ is seen to cause a number of PM-RO sequences. This results from the various manners in which different speakers merge and split the formants characteristic of the terminal /r/. Many of the PM variations created by different speakers are very brief in duration and are correctly termed minor PM's. These short sequences of minor PM's are seen to be equivalent realizations of the /r/. Fortunately, the data contains five realizations from the five speakers. It is apparent that there are not many other ways that the /r/ can be realized. Thus a larger



Numerals in the circles denote the PM type.
Connecting lines are RO's with arbitrary labels
assigned to indicate similarities

Figure 16. PM-RO Sequence for Each Speaker Articulating /Before/.



Connecting RO's are denoted with arbitrary labels to indicate similarity.

Fig. 17. Collapsed PM-RO Sequence for All Utterances of /Before/.

data sample would cluster more speakers on each of the several PM-RO sequences.

It should be noted that some speakers exhibited only two formants where others exhibited three in the same word. This is due to the particular speaker's physiology realizing two formants so close together that our parameterization of the sonagram does not resolve them. A unique feature of the PM-RO representation is that it allows, in some cases, the comparison of two PM sequences when one is composed of three formant PM's and the other of two formant PM's. This results from the fact that the RO_i's highlight the distinctive characteristics of the transitions between PM's and allow the deemphasis of uninteresting information. Hence, two PM sequences can be said to be equivalent in a dynamic sense even though they contain noncomparable static features.

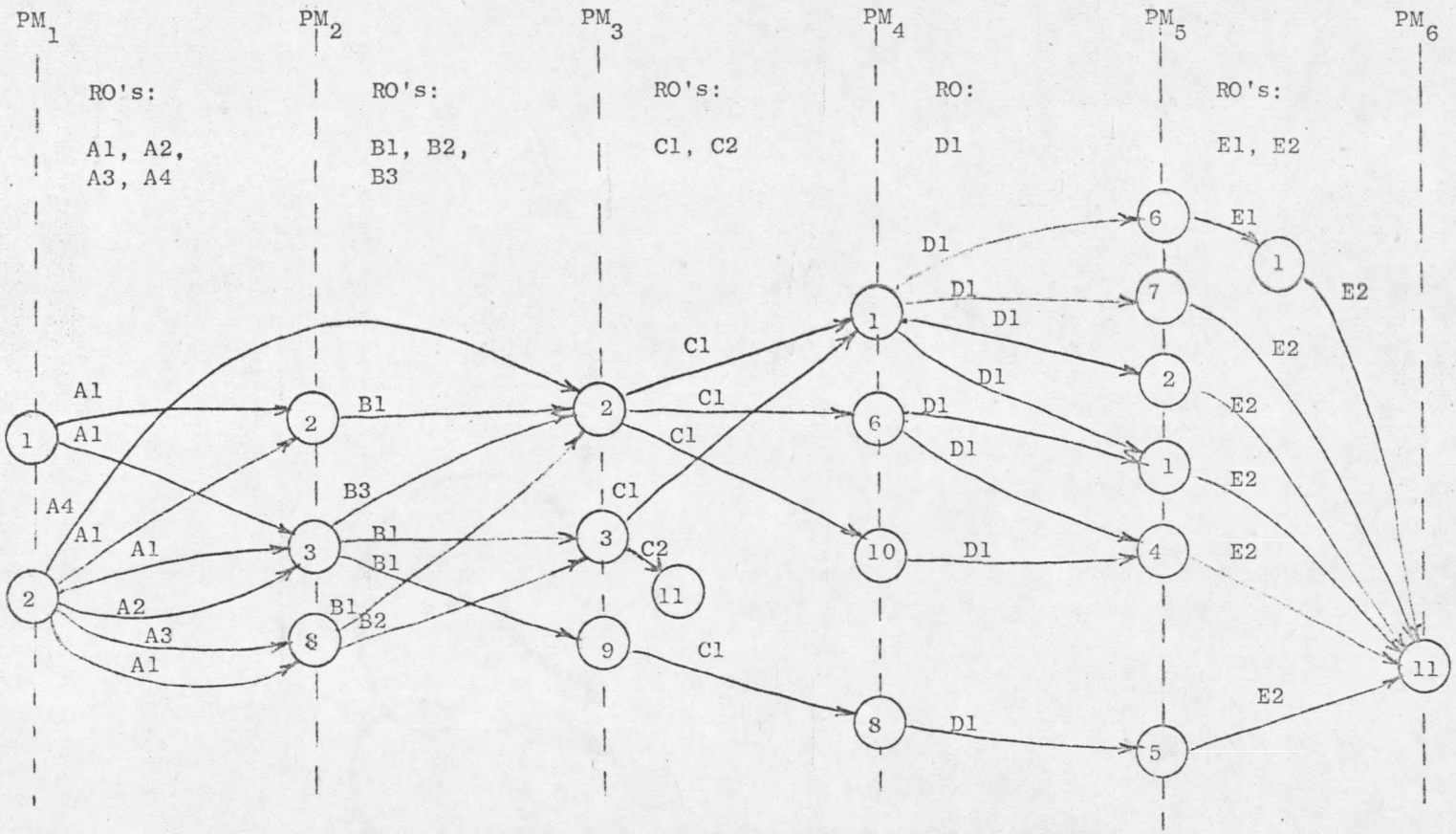
At first glance it appears that a tremendous amount of information has been discarded when the state vector is represented by one of ten PM's. This is not the fact. The information is transferred to the RO. This is not an immediately obvious point, but it may be demonstrated by reconstruction of the state vectors totally from the PM-RO sequence. The RO_i elements are highly correlated in the sense that most transitional features are reflected in several of the ro_j's. There are also a number of cases where RO_i elements are redundant. This appears useful with visual interpretations since it brings to light more easily certain types of transitions.

5.0 DISCUSSION OF PM-RO METHODS AND THEIR RELATION TO THE SPEECH RECOGNITION PROBLEM

The methods described for the segmentation of the speech signal and the related representation as a sequence of Production Modes and Relative Oppositions operate on the P-PHON stratum of our linguistic model. As such, they serve to define the basis of a space in which many of the articulatory movements of an individual speaker are reflected. They provide a very fine-grained p-phonetic representation of the actual speech signal. These p-phonetic units appear to provide sufficiently detailed input for the higher strata. There are numerous advantages to this approach. The initial segmentation procedure is independent of the speaker's voice range and timing. The actual determination of segment boundaries proceeds asynchronously with the input delayed by only three data frames. The formation of the state vector may be done immediately upon the determination of a segment terminal boundary. The determination of the Relative Opposition may proceed one state vector behind the real-time signal. After the Relative Opposition has been computed, the only data storage required is the PM for the i^{th} state vector. Thus, a very large data reduction is effected. Noticeably lacking in our processing techniques is a scheme for the extraction of such supra-segmental features as pitch and stress. In an actual recognition system, the control of supra-segmental feature extraction could be derived from the PM decisions. That is, pitch computations would be performed only where appropriate. The entire processing algorithm is deterministic with information flow only in the forward direction. It is not context sensitive in the sense that future

boundary determinations may not cause previous measurements at the lower levels to be inappropriate. The processing produces a string of PM-RO. Absolute and relative information are contained in both elements of the representation. The PM is a definite statement about the absolute character of the signal segment. The positioning and shape of the peaks within a PM_i is indirectly recorded in the RO_i elements. However, since the RO_i's contain only information about relative change, the recreation of a particular PM may involve the coherent evaluation of information contained in the PM's and RO's on either side of the desired PM_i. This demonstrates that the information of the acoustic signal is actually distributed throughout our representation. Fig. 18 illustrates the existence of short PM-RO sequences that recur in particular speakers' utterances. The fact that several equivalent PM-RO minor sequences represent the realization of the same sound by different speakers is a consequence of our fine-grained measurements. In terms of the strata model, the equating of these minor sequences is done in the p-phonemic rank of the P-PHON stratum. The information that would be passed to the PHON stratum would be in terms of these p-phonemic units. The recognition of the p-phonemic units and sequences thereof on the PHON stratum is essentially the conversion of the phonetic representation to the phoneme representation. The phoneme would be defined as the collection of phonon elements having the same meaning.

The choice of the computations $\{ro_j\}$ that define an RO was not discussed in the previous section because it has important implications with respect to the learning and adaptation capabilities of the system. There are several choices for a philosophy to follow in the selection of the ro_j elements.



The collapsed PM-RO sequence for the second syllable of forward and downward articulated by all five speakers. The RO labels (e.g. A1, A2, A3, A4) denote transitions with nearly the same collection of distinctive changes; consequently, it is possible for the same RO to exist between different PM_i and PM_{i+1}.

Figure 18. Sample of Stable Recurring Sequences of Minor PM-RO.

The method used here was to choose a set $\{ro_j\}$ which characterized every conceivable event in the PM transition. For many PM changes, this means the information in the RO_i is redundant and irrelevant. It also requires the use of the indeterminant result (I) for computations which cannot be performed for a particular PM transition. The rationale for choosing a complete set of $\{ro_j\}$ elements was their usefulness in obtaining information about universal features: those features of each particular utterance that were used by all five speakers. This is accomplished by selecting corresponding PM_i RO_i^A PM_{i+1} ... sequences from each of the speakers uttering the same vocabulary entry (e.g. word A). Further discussion of the information contained in the RO_i is provided by comparing corresponding PM_i RO_i^B PM_{i+1} sequences (where PM_i and PM_{i+1} are the same as those of word A) from each speaker uttering a different vocabulary entry (e.g. word B). The interesting elements (the ro_j) of RO_i^A and RO_i^B are those that are consistent across speakers and may be used to either equate or discriminate A and B. A precise description of sets of these consistent ro_j elements is made more tractable by a cartesian cross product interpretation of the RO_i elements from different utterances.

The general form of a Relative Opposition contained in a vocabulary word A articulated by speaker m, and associated with the transition from PM_i to PM_{i+1} is:

$$RO_i^{A_m} = \left\{ ro_1, ro_2, \ldots ro_j, ro_{j+1}, \ldots ro_{j_{max}} \right\} = \bigtimes_J ro_j$$

$$J = \left\{ 1, 2, \ldots j_{max} \right\}$$

The values which each ro_j may have are blank (blk), indeterminant (I), and numerical (denoted by n with a subscript). For a word A articulated by a set of $M = \{m_1, m_2, \ldots m_{max}\}$ speakers, we divide the elements of relative opposition derived from all speakers (RO_i^A) into three sets: same (s), different (d), and blank (b).

$$RO_i^A = \left(\underbrace{\text{Same}}_{j \epsilon J_s^A} ro_j \right) \left(\underbrace{\text{Different}}_{j \epsilon J_d^A} ro_j \right) \left(\underbrace{\text{Blank}}_{j \epsilon J_b^A} ro_j \right)$$

or, in another form:

$$RO_i^A = (ro_j)_{j \epsilon J_s^A} (ro_j)_{j \epsilon J_d^A} (ro_j)_{j \epsilon J_b^A}$$

$$J_s^A \cup J_d^A \cup J_b^A = J$$

$$J_s^A \cap J_d^A = J_s^A \cap J_b^A = J_d^A \cap J_b^A = \Phi$$

Let m_1 index one speaker's RO_i^A (denoted $RO_i^{A_{m_1}}$) and m_2 index another speaker's RO_i^A (denoted $RO_i^{A_{m_2}}$). The following rules define J_s^A, J_d^A, and J_b^A. We are eliminating the dependence on $\{m_1, m_2, \ldots m_{max}\}$ by applying the rules to pairs m_1, m_2; m_2, m_3; m_3, m_4; ..., until the set is exhausted.

σ. $j_o \epsilon J_s^A$ iff the values for $ro_{j_o}^{m_1}$ and $ro_{j_o}^{m_2}$ satisfy one of the following.

$\sigma 1$. $ro_{j_o}^{m_1} = ro_{j_o}^{m_1} = I$

$\sigma 2$. $ro_{j_o}^{m_1} = n_{m_1}$, $ro_{j_o}^{m_2} = n_{m_2}$

and $\left| n_{m_1} - n_{m_2} \right| \leq \gamma$ where γ is a tolerance constant

δ. $j_o \in J_d^A$ iff the values for $ro_{j_o}^{m_1}$ and $ro_{j_o}^{m_2}$ satisfy one of the following:

$\delta 1$. $ro_{j_o}^{m_1} = I$, $ro_{j_o}^{m_2} = blk$

$\delta 2$. $ro_{j_o}^{m_1} = I$, $ro_{j_o}^{m_2} = n_{m_2}$

$\delta 3$. $ro_{j_o}^{m_1} = n_{m_1}$, $ro_{j_o}^{m_2} = blk$

$\delta 4$. $ro_{j_o}^{m_1} = n_{m_1}$, $ro_{j_o}^{m_2} = n_{m_2}$

and $\left| n_{m_1} - n_{m_2} \right| > \gamma$ where γ is a tolerance constant

β. $j_o \in J_b^A$ iff the values for $ro_{j_o}^{m_1}$ and $ro_{j_o}^{m_2}$ satisfy the following:

β. $ro_{j_o}^{m_1} = ro_{j_o}^{m_2} = blk$

In the case that condition $\sigma 2$ is satisfied

set $$ro_{j_o}^{m_a^*} = \frac{n_{m_1} + n_{m_a}}{2} = n_{m_2}^*$$

The process is continued by applying the rules to $RO_i^{A_{m_3}}$ with the elements of $RO_i^{A_{m_2}}$ replaced by $ro_{j_o}^{m_2^*}$ as required. In any cases that the $\sigma 2$ condition is satisfied, replace the appropriate elements with the weighted average.

$$ro_{j_o}^{m_3^*} = \frac{2\, n_{m_2} + n_{m_3}}{3} = n_{m_3}^*$$

A short numerical example of this process is included as Appendix C.

For two vocabulary words, A and B, this leads to six index sets.

$$J_s^A, \quad J_d^A, \quad J_b^A$$

$$J_s^B, \quad J_d^B, \quad J_b^B$$

The product of the J_s^A and J_s^B sets is:

$$J_s^{AB} = \left\{ j : j \in J_s^A \cap J_s^B \text{ and satisfying condition } \sigma^* \right\} \quad \underline{\text{Type I}}$$

i.e. $RO_{i_s}^{AB} = \bigtimes_{j \in J_s^{AB}} ro_j$ is the set of same (in value) ro_j across speakers and utterances

Now define two additional index sets:

$$J_{II}^A = J_s^A - J_s^{AB}$$

$$J_{II}^B = J_s^B - J_s^{AB}$$

It is true only that:

$$\phi \leq J_{II}^A \cap J_{II}^B$$

because of having to satisfy condition σ.

---

* comparing $ro_{j_o}^A$ and $ro_{j_o}^B$ instead of $ro_{j_o}^{m_1}$ and $ro_{j_o}^{m_2}$

J^A_{II} and J^B_{II} index the discriminant sets:

$$RO^A_{i_{II}} = \bigtimes_{j \in J^A_{II}} ro_j \qquad RO^B_{i_{II}} = \bigtimes_{j \in J^B_{II}} ro_j \qquad \underline{\text{Type II}}$$

The remaining elements, ro_j, from RO^A_i and RO^B_i are those which were not consistent across either speakers or vocabulary words.

$$RO^{AB}_{i_r} = \bigtimes_{j \in J^{AB}_r} ro_j \qquad \underline{\text{Type III}}$$

$$J^{AB}_r = \left(\overline{J^A_s \bigcup J^B_s}\right)$$

The significance of the Type I, II, and III sets is that they separate the acoustic events in the speech signal into classes that reflect the information content and the consistency across speakers. In the particular case where PM^A_i RO^A_i PM^A_{i+1} must be differentiated from PM^B_i RO^B_i PM^B_{i+1}, they focus attention on the distinctive characteristics of A and B transitions. The ro_j of the Type I sets are always the same for A and B for all speakers. The ro_j of the TypeIII sets are not consistent across all speakers for either utterance A or B. The ro_j in the Type II sets are those that were the same for all speakers articulating either A or B but not both A and B with the same value. The precise definition of Special Type II sets will simplify further discussion of their properties.

The Type II sets previously defined, J^A_{II} and J^B_{II}, may be expressed as unions of two special sets.

$$J^A_{II} = {}^{cf}J^{AB}_{II} \cup {}^{cs}J^A_{II}$$

$$J^B_{II} = {}^{cf}J^{AB}_{II} \cup {}^{cs}J^B_{II}$$

where the special sets are defined as follows:

$${}^{cf}J^{AB}_{II} \stackrel{d}{=} \left(J^A_s \cap J^B_s \right) - J^{AB}_s$$

$${}^{cs}J^A_{II} \stackrel{d}{=} J^A_s \cap \left(J^B_d \cup J^B_b \right)$$

$${}^{cs}J^B_{II} \stackrel{d}{=} J^B_s \cap \left(J^A_d \cup J^A_b \right)$$

The set of ro_j elements in the new discriminant set

$${}^{cf}RO^{AB}_{i_{II}} = \bigtimes_{j_1 \in {}^{cf}J^{AB}_{II}} ro_{j_1}$$

separate utterances A and B for all speakers. The cf superscript draws attention to the context-free power of this set of ro_j. The other special Type II sets:

$${}^{cs}RO^A_{i_{II}} = \bigtimes_{j_1 \in {}^{cs}J^A_{II}} ro_{j_1} \quad \text{and} \quad {}^{cs}RO^B_{i_{II}} = \bigtimes_{j_1 \in J^B_{II}} ro_j$$

are context sensitive and are not guaranteed to provide an absolute differentiation of A and B for all speakers. This situation arises when an ro_j element is consistent across speakers for one utterance (e.g. A) and varies across speakers for another utterance (e.g. B). Consider a particular ro_{j_0} that belongs to ${}^{cs}RO^{A}_{i_{II}}$ and that has a consistent numerical value $n_0 \pm \gamma/2$ for word A across speakers. The occurrence of $\left|ro_{j_0} - n_0\right| > \gamma$ is sufficient to identify not A. However, within the context of the total PM-RO sequence it may result that $ro_{j_0} \in {}^{cs}RO^{A}_{i_{II}}$ has absolute discriminant power even when $\left|ro_{j_0} - n_0\right| \leq \gamma$. In subsequent discussion, we will refer to a Type II_T set that is composed of a Type II_{cf} set and a Type II_{cs} set.

If we consider the case of training a machine constructed with a predetermined set of $\{ro_j\}$, then the machine has a maximum capacity for that set of $\{ro_j\}$. In the interest of efficiency the minimum set providing for separation of the various responses would be chosen. More than likely this would be an ad hoc design in which sets of $\{ro_j\}$ were selected and tested until sufficient separation of the input stimuli and the responses was achieved. This is measured by the richness of the Type II_{cf} and Type II_{cs} sets. The characteristics of a directed search are provided by the possibility of computing only a restricted set of ro_j at a particular PM_i which are known to possess useful information content (i.e. those specified by the pertinent Type II_T set). Additions to the vocabulary would require the evaluation of the Type II_{cf} and Type II_{cs} sets. If the addition of vocabulary elements meant the disappearance of the unique ro_j elements of the Type II_T sets, then a more

detailed or a more complete set of $\{ro_j\}$ would be required.

There are two system objectives for which it might be desirable to automate the incremental generation of ro_j measurements. The evaluation of the information-bearing content of each $\{ro_j\}$ for a particular vocabulary could be determined from the distribution of the membership of each ro_j in the Type II_T discriminant sets. The method of approach would parallel the type of automatic feature selection algorithms implemented by Urh [20]. The method is to provide a set of primitive measurements (the ro_j elements) and a set of representative data. For the evaluation of the hierarchical aspects of the features, one presents the total vocabulary and evaluates the usefulness of each ro_j computation by its appearance in the discriminant sets. The $\{ro_j\}$ measurements are then simply reordered with respect to their effect in producing correct responses. The number of ro_j is only as large as that required to produce separation of the desired responses. The number of computations is large, but the computations are straightforward. Each ro_{j_i} must be applied to each input and its results evaluated by a separability criterion.

An attempt at imitating human initial language acquisition could be made in a similar manner, but the sequence of ro_j evaluations is more complicated. The procedure is to provide a set of primitive ro_j elements and then allow for their incremental selection while presenting suitable data in a same-different ordering. Suitable data consists of samples:

$$S = \left\{ s_1^1 , s_2^1 ; s_1^2 , s_2^2 ; \ldots ; s_1^k , s_2^k \right\}$$

where the superscript denotes the input category and the 1,2 subscripts indicate different samples within the category. The same-different ordering results in a special presentation of the training inputs.

Input Cycle	Total Training Samples	New Training Samples	Comparison Operation
1	---	s_1^1 s_2^1	SAME
2	s_1^1 s_2^1	s_1^2	DIFFERENT
3	s_1^1 s_2^1 s_1^2	s_2^2	SAME
4	s_1^1 s_2^1 s_1^2 s_2^2	s_1^3	DIFFERENT
5	s_1^1 s_2^1 s_1^2 s_2^2 s_1^3	s_2^3	SAME
...			
(2k-1)	s_1^1 s_2^1 ... s_1^k	s_2^k	SAME

On input cycle 1, the first ro_j in the list of primitive calculations that provides the same measure on the two same inputs is selected and stored as the description of them. On input cycle 2, an ro_j (possibly the one selected in cycle 1) is found that separates s_1^2 and (s_1^1 , s_2^1). Input cycle 3 finds an ro_j that places s_1^2 and s_2^2 in the same category. The ro_j that place samples in the same categories together and the different categories become the pattern's descriptor. Note that the new descriptor (ro_j calculation) must be added to the list of descriptions of the previous patterns on the conclusions of every other input cycle. The result is that each time a new ro_j is added to the RO_i list

and found to be useful, it must be applied to every previous sample and added to all lists of descriptors. A further operation is that of cycling through the set of descriptors and eliminating those which were previously useful but are no longer useful in the expanded vocabulary.

The design of an automated system for the extraction of the semantic information based on the PM-RO sequences could easily provide information on speaker identity. It would require the addition of a storage system to record the applicable ro_j of the Type III sets that each speaker exhibited. The exact values of the Type III ro_j elements for each speaker could be used as a speaker signature. The statistics of these effects could easily be incorporated into an open-ended speaker identification system, i.e. "open-ended" implying that the speaker is recognized as one of a number of previously known speakers or is unknown.

6.0 CONCLUSIONS

A solution, in concept, to the problem of segmenting continuous speech is provided by defining homogeneous epochs of the acoustic speech signal suitable as the basis of an automatic system for the extraction of linguistically encoded information. The methods and algorithms developed for the processing of the speech signal as recorded in the sonagram are intended only as examples of the design approach. It appears that the inaccuracy and lack of resolution of the sonagrams is compensated for by their high information content as visual displays of the total utterance. Unfortunately, they do not provide the supra-segmental features of stress and pitch. It cannot be overemphasized that the concepts of predictive segmentation, state vector, Production Mode and Relative Opposition are not in any way tied to the equipment. Given any reasonably accurate continuous record of the speech signal short-term spectrum, algorithms to realize a P-PHON stratum representation could be easily spectrified. It would be wise to include provisions for the extraction of pitch contours and energy of the appropriate segments.

Using the design approach expounded in this paper, it should be possible to realize an automatic system for the extraction of semantic content. This system should be capable of providing a crude graphical translation of a large class of English sentences made up from a fixed vocabulary. The major constraints on the set of utterances will undoubtedly be sentence structure; e.g., utterances may be restricted to simple declarative statements and questions. The primary vocabulary constraint will

probably be the total number of words. The graphical translation should be in the form of rough-draft English readable by the average speaker.

The major problem to be overcome, however, is not the number of items which must be stored in a machine memory, but rather the number of items to which the input must be compared at each point in the decision process. If each minute segment of the input must be compared with ten thousand stored descriptions, the process becomes highly inefficient. If some method is available that allows the input to be compared with a subset (say ten out of ten thousand) of the stored descriptions, the process becomes efficient and possible in real time. Use of our PM-RO representation at the P-PHON stratum provides this directed search through the Type II discriminant sets. The homogeneous segments of the input signal are converted to PM-RO representation as p-phonetic units which need be compared only to the sequence of minor states stored as p-phonemic units within the P-PHON stratum. Once the p-phonemic element has been identified, all further analysis is carried out in the PHON stratum in terms of the context of phonons. Most likely these phonons will be similar to the phonons of any classical phonology. The limited search is again realized on the PHON stratum in the equivalencing of phonon sequences with units of phonemic identity. The supra-segmental features need either to be carried up through the various strata as augmentations at each level or entered at the MORPH and LEX strata. This is appropriate since the supra-segmental system of features (grammar) is directed at the MORPH stratum.

The use of context is achieved on each stratum. Context on the P-PHON stratum is dictated primarily by the dynamics of the patterned articulatory movements and shows up in the Type II_{cs} discriminant sets. Context within the PHON stratum is normally called phonotactics and involves the acceptable sound alternations that words undergo as they are concatenated to form a continuous utterance. MORPH stratum context is concerned with alternations in the phonemic composition of words and phrases. The context constraints imposed by the grammar of the language are first encountered in the MORPH stratum and completed in the LEX stratum. From the data presented herein, it does appear that tiepoints, major Production Mode oppositions, do exist and recur in continuous speech. (e.g. The tiepoints are numerous PM_i-RO_i, like voiceless fricatives, that have well-populated Type II_{cf} discriminant sets.) Linguistic theory is well developed concerning the tiepoints of context at the higher levels. These are the supra-segmental features of pause, stress, and pitch contours.

By far the most important characteristic of this decision structure for natural language is the absence of the requirement for any closeness measure defined over the absolute characteristics of sound units (or phonemes). A closeness measure does exist, but it is not wholly in terms of absolute features. The distribution of information over all the elements of a PM-RO sequence allows the computation of a distance measure that is not bound to the comparison of two elements alone. Any likeness between acoustic signals is computed in terms of the PM-RO sequences between major tiepoints and can make use of the elements of each Type II_{cs} discriminant

set. This measure is not bound to structuring the phonetic elements of speech strictly by their acoustic character. It provides a basis for closeness which is very similar to human perceptual behavior. Human perceptual errors may be directly related to the sound unit's relationship in the sequence of sound units. Examples of human perceptual errors may be found in everyday occurrences. For example, many cases of perceptual confusion transcend normal boundaries established between sounds by place of articulation or manner of articulation; incorrect perception of a /t/ in place of an /n/ in a vowel/consonant/vowel environment is a good example of this.

The data for this investigation was chosen for its representation of a wide variety of sounds produced in context by five speakers. Consequently, it requires only the initial portions of all ten words to uniquely identify the words regardless of speaker. The allowable conclusions are restricted to the manner of speech production used by five speakers to achieve the same perceived utterances. The representation of their utterances in the PM-RO sequences clearly indicates the existence of certain universally controlled features in the production of the same words. These common distinctive features are apparent in the ro_j elements of the Type II_T sets representing the transitions between PM's. The five speakers all added to their speech individual features (the ro_j elements thata result in the Type III sets) that are continuous across the PM boundaries. Since the RO representation does not attach significance for discrimination to these continuous features, it appears to be a very useful representation. It results in a representation of

speech that distributes the information-carrying load between the absolute character of homogeneous segments and the transitions between these homogeneous segments. The Relative Opposition provides a set of pointers that illuminate the features carrying the major information in any Production Mode. They provide a context-free directed search mechanism at all linguistic levels from the acoustic signal to the sentence.

The processing algorithms that have been demonstrated on the sonagrams should not necessarily be realized in an operational system on a power spectrum representation of the speech signal. They might just as well, and more practically, be realized by direct processing of the acoustic time series. A model that represents the homogeneous segments of the speech signal as a set of time-varying statistical estimates seems appropriate. Predictive segmentation could easily be realized when it appears that the present statistical model is no longer accurately predicting the incoming signal. At the end of a segment boundary, the time series model contains a time history of the segment and could be converted to an equivalent state vector. The Relative Oppositions could be realized as the comparison of generation model parameters between homogeneous segments. This would provide abstract parameters that could represent the continuous changes of a formant just as completely as do a series of short-term power spectra. The primary difficulty lies in relating the highly abstract system parameters to the speaker's articulation. The choice of sonagrams (power spectrum

representation) in this investigation allowed the use of many years of experience by other investigators with the relationship between articulatory dynamics and spectrographic records. Although the parameters used in this investigation were somewhat abstract, they can still be related to previous formant analyses of speech.

APPENDIX

APPENDIX A - LITERATURE CITED

[1] J. L. Flanagan, "Speech Analysis Synthesis and Perception," Academic Press, Inc., New York, New York, 1962.

[2] K. L. Pike, "Phonemics: A Technique for Reducing Languages to Writing," University of Michigan Press, Ann Arbor, Michigan, 1947.

[3] S. M. Lamb, "Prolegomena to a Theory of Phonology," Language, Vol. 42, 1966.

[4] N. Lindgren, "Machine Recognition of Human Language," (A three-part review of automatic speech recognition, theoretical models of speech perception and language, and cursive script recognition), IEEE Spectrum, March-May 1965.

[5] D. Tufts et al, "Some Results in Automatic Speech Segmentation Using Wide-Band Filtering and Subtraction," JASA, Vol. 37, 1965.

[6] J. B. Thomas, "The Significance of the Second Formant in Speech Intelligibility," TR-10, University of Illinois, Urbana, Illinois, July, 1966.

[7] T. B. Martin et al, "Speech Recognition by Feature-Abstraction Techniques," Tech. Doc. Report, No. AL TDR 64-176. AF Avionics Lab., Wright-Patterson AFB, Ohio, August 1964.

[8] S. E. G. Öhman, "Coarticulation in VCV Utterances: Spectrographic Measurements," JASA, Vol. 37, 1966.

[9] R. K. Potter, G. A. Kopp, and H. C. Green, "Visible Speech," D. Van Nostrand Co., Inc., New York, New York, 1947.

[10] A. M. Libereman, "Some Results of Research on Speech Perception," JASA, Vol. 28, 1957.

[11] G. A. Miller and P. E. Nicely, "An Analysis of Perceptual Confusion Among Some English Consonants," JASA, Vol. 27, 1955.

[12] S. Singh and J. W. Black, "Study of Twenty-Six Intervocalic Consonants as Spoken and Recognized by Four Language Groups," JASA, Vol. 39, 1966.

[13] W. A. Wickelgren, "Distinctive Features and Errors in Short-Term Memory for English Consonants," JASA, Vol. 39, 1966.

[14] K. M. Sayre, "Recognition: A Study in the Philosophy of Artificial Intelligence," Notre Dame,Press, South Bend, Indiana, 1965.

[15] J. Gazdag, "A Method of Decoding Speech," Report No. 9, Engineering Experiment Station, University of Illinois, Urbana, Illinois, 1966.

[16] D. G. Bobrow and D. H. Klatt, "A Limited Speech Recognition System," Final Report, NAS 12-138, NASA, May 1968.

[17] N. R. French and J. C. Steinberg, "Factors Governing the Intelligibility of Speech Sounds," JASA, Vol. 19, 1947.

[18] D. R. Reddy, "An Approach to Computer Speech Recognition by Direct Analysis of the Speech Wave," TR CS49, Computer Science Dept., Stanford University, Stanford, California, 1966.

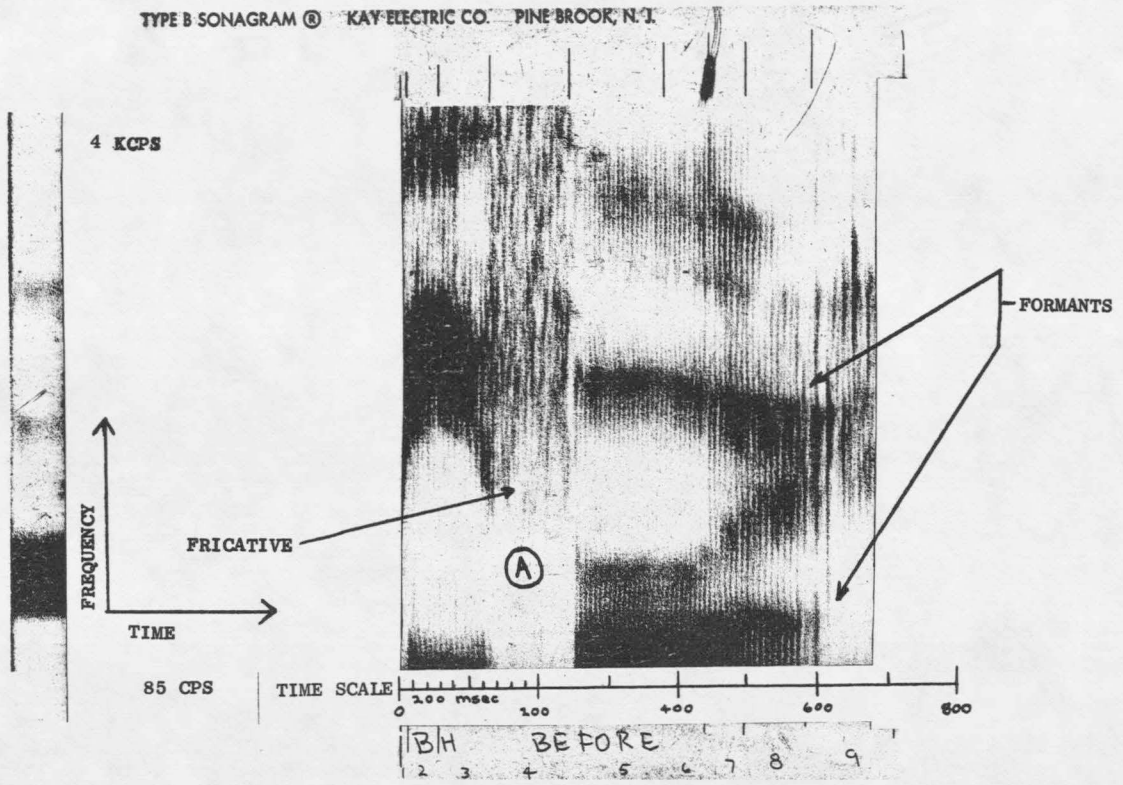
[19] D. R. Reddy and A. E. Robinson, "Phoneme to Grapheme Translation of English," IEEE Transaction on Audio and Electroacoustics, Vol. AU-16, 1968.

[20] L. Uhr, "Pattern Recognition," edited by L. Uhr, John Wiley and Sons, Inc., New York, New York, 1966.

APPENDIX B

Two Speakers' Sonagrams, State Vectors, RO's,

and PM-RO Sequences for the Utterance /Before/.



Sample Sonagram of /Before/.

TIME		STATE	DATA VECTORS
00	000000000122333333321111111222111	1	/BEFORE/ BH
01	000000000122333333321111111222111		
02	320000000000133333332100000122221	2	
05	320000000000002333333310000122221		
06	320000000000002333333310000122221	3	
13	320000000000123333321000000000111		
14	000000000022222222222111111111111N	4	
24	000000000022222222222111111111111N		
25	333222000000001232000000011110000	5	
38	333222000000001232000000011110000		
39	333222000000001232000000011110000	6	
44	333222000000012321000111111000000		
45	333222200000123210001111110000000	7	
50	333211122101232100011111100000000		
51	333200022212233100000000000000000	8	
60	333100012222222100000000000000000		
61	000000001222222222211111000000000N	9	
70	000000001222222222211111000000000N		

```
STATE
1    0/1/43/43/01///43/ 0/2D/2D//10/ 0//32/ 0///

2    0/3/38/37/03/// 5/ 0/1D/1D// 1/ 0// 2/ 0///
                  ///25/-1/BP/BP//13/ 2//22/ 1///
                  /// 8/ 0/BP/BP//28/ 0//32/ 0///

3    0/3/37/29/07/// 5/ 0/1D/1D// 1/ 0// 2/ 0///
                  ///24/-3/BP/BP//15/-2//23/-2///
                  /// 8/-5/BP/BP//28/ 2//32/ 0///

4    1/1/34/34/10///34/ 0/1D/1D//11/ 0//32/ 0///

5    0/3/27/27/13///15/ 0/1D/1D// 1/ 0// 6/ 0///
                  /// 8/ 0/SP/SP//15/ 0//18/ 0///
                  /// 4/ 0/BP/BP//26/ 0//29/ 0///

6    0/3/27/30/05///15/ 0/1D/1D// 1/ 0// 6/ 0///
                  /// 8/ 1/SP/SP//15/-1//18/ 0///
                  /// 4/ 2/BP/BP//26/-4//27/-1///

7    0/3/32/34/05///17/ 2/BP/BP// 1/ 0// 6/ 3///
                  /// 9/ 0/SP/SP//13/-1//17/-1///
                  /// 6/ 0/BP/BP//21/-1//26/-1///

8    0/2/29/26/09///11/-1/BP/BP// 1/ 0// 4/ 0///
                  ///18/-2/1U/BP// 8/+1//16/ 0///

9    1/1/26/26/09///26/ 0/1D/1D// 9/ 0//24/ 0///
```

PM_i S

PM_{i+1} 6

PM_i SEQ NO 0

SAMPLE WORD 5-BH

Duration Ratio 5/1

Power Ratio 5/43

Power Distribution 5 ____ ____ / 100 ____ ____

Merge; Split 0 0 Boundary Revisions (Including Those Renamed in PM_{i+1})

DESTROYED	L1		U1		L2		U2		L3		U3	
CREATED	L1		U1		L2		U2		L3		U3	

PK - 1 Information

Abrupt Initiation (Power) 100

Abrupt Termination (Power) 0

Abrupt Boundary Discontinuities — LB Step I — UB Step I

Shapes I/2D — Slopes I/0 — Slopes I/0

PK - 2 Information

Abrupt Initiation (Power) ____

Abrupt Termination (Power) ____

Abrupt Boundary Discontinuities — LB Step ____ — UB Step ____

Shapes ____ — Slopes ____ — Slopes ____

PK - 3 Information

Abrupt Initiation (Power) ____

Abrupt Termination (Power) ____

Abrupt Boundary Discontinuities — LB Step ____ — UB Step ____

Shapes ____ — Slopes ____ — Slopes ____

PM_i 6 PM_i SEQ NO 1

PM_{i+1} 3 SAMPLE WORD 5-BH

Duration Ratio 1/3

Power Ratio 43/38

Power Distribution 100 ___ ___ / 13 66 21

Merge; Split 0 0 Boundary Revisions (Including Those Renamed in PM_{i+1})

DESTROYED	L1		U1		L2		U2		L3		U3	
CREATED	L1		U1		L2		U2		L3		U3	

PK - 1 Information

Abrupt Initiation (Power) 13

Abrupt Termination (Power) 0

Abrupt Boundary Discontinuities LB Step I UB Step I

Shapes I/1D Slopes I/0 Slopes I/0

PK - 2 Information

Abrupt Initiation (Power) 0

Abrupt Termination (Power) 0

Abrupt Boundary Discontinuities LB Step +3 UB Step I

Shapes I/BP Slopes 0/+.6 Slopes I/+3

PK - 3 Information

Abrupt Initiation (Power) 0

Abrupt Termination (Power) 0

Abrupt Boundary Discontinuities LB Step I UB Step 0

Shapes I/BP Slopes I/0 Slopes 0/0

PM_i 3 PM_i SEQ NO 2

PM_{i+1} 3 SAMPLE WORD S-BH

Duration Ratio 3/7

Power Ratio 37/37

Power Distribution 14 65 21 / 14 65 21

Merge; Split 0 0 Boundary Revisions (Including Those Renamed in PM_{i+1})

DESTROYED	L1		U1		L2		U2		L3		U3	
CREATED	L1		U1		L2		U2		L3		U3	

PK - 1 Information

Abrupt Initiation (Power) 0

Abrupt Termination (Power) 0

Abrupt Boundary Discontinuities LB Step 0 UB Step 0

Shapes 1D/1D Slopes 0/0 Slopes 0/0

PK - 2 Information

Abrupt Initiation (Power) 0

Abrupt Termination (Power) 0

Abrupt Boundary Discontinuities LB Step 0 UB Step 0

Shapes BP/BP Slopes +.66/-.28 Slopes +.33/-.28

PK - 3 Information

Abrupt Initiation (Power) 0

Abrupt Termination (Power) 0

Abrupt Boundary Discontinuities LB Step 0 UB Step 0

Shapes BP/BP Slopes 0/+.28 Slopes 0/0

PM_i 3 PM_i SEQ NO 3

PM_{i+1} 4 SAMPLE WORD S-BH

Duration Ratio 7/10

Power Ratio 29/34

Power Distribution 17 73 10 / 100

Merge; Split 0 0 Boundary Revisions (Including Those Renamed in PM_{i+1})

DESTROYED L1 ____ U1 ____ L2 ____ U2 ____ L3 ____ U3 ____

CREATED L1 ____ U1 ____ L2 ____ U2 ____ L3 ____ U3 ____

PK - 1 Information

Abrupt Initiation (Power) 0

Abrupt Termination (Power) 17

Abrupt Boundary Discontinuities LB Step I UB Step I

Shapes 1D/I Slopes 0/I Slopes 0/I

PK - 2 Information

Abrupt Initiation (Power) 0

Abrupt Termination (Power) 0

Abrupt Boundary Discontinuities LB Step -2 UB Step I

Shapes BP/I Slopes -.28/0 Slopes +.28/I

PK - 3 Information

Abrupt Initiation (Power) 0

Abrupt Termination (Power) 0

Abrupt Boundary Discontinuities LB Step I UB Step 0

Shapes BP/I Slopes +.28/I Slopes 0/0

PM_i 4 PM_i SEQ NO 4

PM_{i+1} 3 SAMPLE WORD 5-B14

Duration Ratio 10/13

Power Ratio 34/27

Power Distribution 100 ___ ___ / 55 30 15

Merge; Split 0 0 Boundary Revisions (Including Those Renamed in PM_{i+1})

DESTROYED L1 ___ U1 ___ L2 ___ U2 ___ L3 ___ U3 ___

CREATED L1 ___ U1 ___ L2 ___ U2 ___ L3 ___ U3 ___

PK - 1 Information

Abrupt Initiation (Power) 55

Abrupt Termination (Power) 0

Abrupt Boundary Discontinuities LB Step I UB Step I

Shapes I/1D Slopes I/0 Slopes I/0

PK - 2 Information

Abrupt Initiation (Power) 0

Abrupt Termination (Power) 0

Abrupt Boundary Discontinuities LB Step +4 UB Step I

Shapes I/SP Slopes 0/0 Slopes I/0

PK - 3 Information

Abrupt Initiation (Power) 0

Abrupt Termination (Power) 0

Abrupt Boundary Discontinuities LB Step I UB Step -3

Shapes I/BP Slopes I/0 Slopes 0/0

PM_i 3 PM_i SEQ NO 5

PM_{i+1} 3 SAMPLE WORD 5-BH

Duration Ratio 13/5

Power Ratio 27/27

Power Distribution 55 30 15 / 55 30 15

Merge; Split 00 Boundary Revisions (Including Those Renamed in PM_{i+1})

DESTROYED	L1		U1		L2		U2		L3		U3	
CREATED	L1		U1		L2		U2		L3		U3	

PK - 1 Information

Abrupt Initiation (Power) 0

Abrupt Termination (Power) 0

Abrupt Boundary Discontinuities LB Step 0 UB Step 0

Shapes 1D/1D Slopes 0/0 Slopes 0/0

PK - 2 Information

Abrupt Initiation (Power) 0

Abrupt Termination (Power) 0

Abrupt Boundary Discontinuities LB Step 0 UB Step 0

Shapes SP/SP Slopes 0/-.2 Slopes 0/0

PK - 3 Information

Abrupt Initiation (Power) 0

Abrupt Termination (Power) 0

Abrupt Boundary Discontinuities LB Step 0 UB Step -2

Shapes BP/BP Slopes 0/-.8 Slopes 0/-.2

PM_i 3 PM_i SEQ NO 6

PM_{i+1} 3 SAMPLE WORD S-BH

Duration Ratio 5/5

Power Ratio 30/32

Power Distribution 50 30 20 / 53 28 19

Merge; Split 0 0 Boundary Revisions (Including Those Renamed in PM_{i+1})

DESTROYED L1 ____ U1 ____ L2 ____ U2 ____ L3 ____ U3 ____

CREATED L1 ____ U1 ____ L2 ____ U2 ____ L3 ____ U3 ____

PK - 1 Information

Abrupt Initiation (Power) 0

Abrupt Termination (Power) 0

Abrupt Boundary Discontinuities LB Step 0 UB Step 0

Shapes ID/BP Slopes 0/0 Slopes 0/+.6

PK - 2 Information

Abrupt Initiation (Power) 0

Abrupt Termination (Power) 0

Abrupt Boundary Discontinuities LB Step 0 UB Step 0

Shapes SP/SP Slopes -.2/-.2 Slopes -.2/-.2

PK - 3 Information

Abrupt Initiation (Power) 0

Abrupt Termination (Power) 0

Abrupt Boundary Discontinuities LB Step 0 UB Step 0

Shapes BP/BP Slopes -.8/-.2 Slopes -.2/-.2

PM_i 3 PM_i SEQ NO 7

PM_{i+1} 2 SAMPLE WORD S-BH

Duration Ratio 5/9

Power Ratio 34/29

Power Distribution 56 27 17 / 38 62

Merge; Split 1 1 Boundary Revisions (Including Those Renamed in PM_{i+1})

	L1	U1	L2	U2	L3	U3
DESTROYED	0	1	1	0	0	0
CREATED	0	1	1	0	0	0

PK - 1 Information

Abrupt Initiation (Power) 0

Abrupt Termination (Power) 0

Abrupt Boundary Discontinuities LB Step 0 UB Step I

Shapes I/BP Slopes 0/0 Slopes I/0

PK - 2 Information

Abrupt Initiation (Power) 0

Abrupt Termination (Power) 0

Abrupt Boundary Discontinuities LB Step I UB Step 0

Shapes I/IU Slopes I/+.11 Slopes -.2/0

PK - 3 Information

Abrupt Initiation (Power) 0

Abrupt Termination (Power) 17

Abrupt Boundary Discontinuities LB Step I UB Step I

Shapes BP/I Slopes -.2/I Slopes -.2/I

PM_i 2 PM_i SEQ NO 8

PM_{i+1} 4 SAMPLE WORD S-BH

Duration Ratio 9/9

Power Ratio 26/26

Power Distribution 38 62 ___ / 100 ___ ___

Merge; Split 0 0 Boundary Revisions (Including Those Renamed in PM_{i+1})

DESTROYED	L1		U1		L2		U2		L3		U3	
CREATED	L1		U1		L2		U2		L3		U3	

PK - 1 Information

Abrupt Initiation (Power) 0

Abrupt Termination (Power) 38

Abrupt Boundary Discontinuities LB Step I UB Step I

Shapes BP/I Slopes 0/I Slopes 0/I

PK - 2 Information

Abrupt Initiation (Power) 0

Abrupt Termination (Power) 0

Abrupt Boundary Discontinuities LB Step 0 UB Step +8

Shapes BP/ID Slopes +2/0 Slopes 0/0

PK - 3 Information

Abrupt Initiation (Power) ___

Abrupt Termination (Power) ___

Abrupt Boundary Discontinuities LB Step ___ UB Step ___

Shapes ___ Slopes ___ Slopes ___

PM_i 4 PM_i SEQ NO 9

PM_{i+1} 5 SAMPLE WORD S-BH

Duration Ratio 9/5

Power Ratio 26/5

Power Distribution 100 ____ ____ / 5 ____ ____

Merge; Split 00 Boundary Revisions (Including Those Renamed in PM_{i+1})

DESTROYED	L1		U1		L2		U2		L3		U3	
CREATED	L1		U1		L2		U2		L3		U3	

PK - 1 Information

Abrupt Initiation (Power) 0

Abrupt Termination (Power) 100

Abrupt Boundary Discontinuities LB Step I UB Step I

Shapes 1D/I Slopes 0/I Slopes 0/I

PK - 2 Information

Abrupt Initiation (Power) ____

Abrupt Termination (Power) ____

Abrupt Boundary Discontinuities LB Step ____ UB Step ____

Shapes ____ Slopes ____ Slopes ____

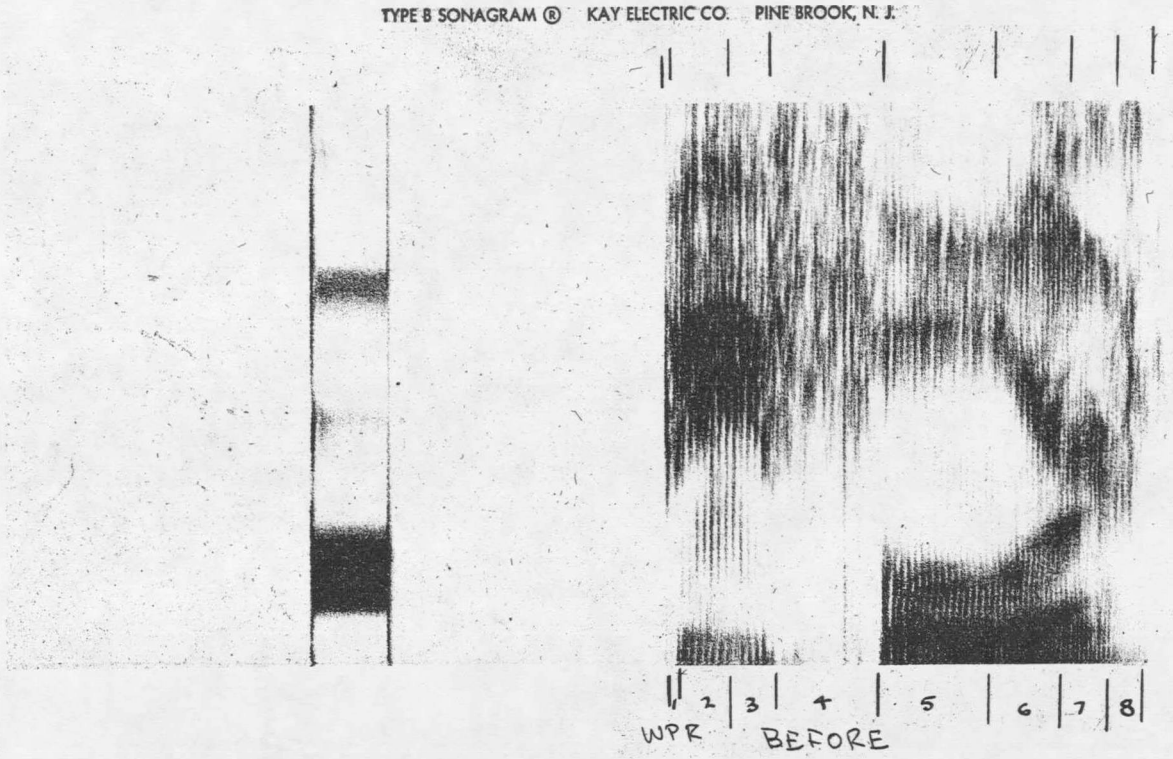
PK - 3 Information

Abrupt Initiation (Power) ____

Abrupt Termination (Power) ____

Abrupt Boundary Discontinuities LB Step ____ UB Step ____

Shapes ____ Slopes ____ Slopes ____


TYPE 8 SONAGRAM ®
KAY ELECTRIC CO.
PINE BROOK, N. J.
1
2
3
4
5
6
7
8
WPR
BEFORE

TIME		STATE	DATA VECTORS
00	000000000012222222222111100000000	1	/BEFORE/WPR
01	000000000012222222222111100000000		
02	320000000012223333332222111100000	2	
08	320000000000012333333322222211000		
09	320000000000012333333322222211000	3	
14	320000000001233333322221111111100		
15	000000000000122222222222222111111N	4	
29	000000000000122222222222222111111N		
30	333321000000000112321000121000000	5	
46	333321000000000112321011110000000		
47	333321000000000112321001111000000	6	
56	333322331012332100000001111000000		
57	333200122222221000000111100000000	7	
62	333000001222221000001111000000000		
63	000000001111111111111000000000000N	8	
68	000000001111111111111000000000000N		

```
STATE
1    0/1/25/25/01///25/ 0/BP/BP//11/ 0//25/ 0///

2    0/2/42/43/06/// 5/ 0/1D/1D// 1/ 0// 2/ 0///
                ///37/ 1/ C/1D//11/ 2//28/ 1///

3    0/2/43/41/05/// 5/ 0/1D/1D// 1/ 0// 2/ 0///
                ///38/-2/1D/1D//13/-1//29/-1///

4    1/1/34/34/14///34/ 0/BP/BP//13/ 0//32/ 0///

5    0/3/30/29/16///15/ 0/1D/1D// 1/ 0// 6/ 0///
                ///10/ 0/SP/SP//16/ 0//21/ 0///
                /// 5/-1/SP/BP//25/-1//27/-1///

6    0/3/29/39/09///15/ 8/1D/BP// 1/ 0// 6/ 3///
                ///10/ 2/SP/SP//16/-5//21/-5///
                /// 4/ 0/BP/BP//24/ 0//26/ 0///

7    0/3/31/24/05///11/-2/BP/BP// 1/ 0// 4/-1///
                ///16/-4/BP/BP// 7/ 2//15/ 0///
                /// 4/-1/BP/BP//24/-1//26/-2///

8    1/1/13/13/05///13/ 0/BP/BP// 9/ 0//21/ 0///
```

PM_i 5 PM_i SEQ NO 0

PM_{i+1} 6 SAMPLE WORD 5-WR

Duration Ratio 5/1

Power Ratio 5/25

Power Distribution 5 ____ ____ / 100 ____ ____

Merge; Split 00 Boundary Revisions (Including Those Renamed in PM_{i+1})

DESTROYED L1 ____ U1 ____ L2 ____ U2 ____ L3 ____ U3 ____

CREATED L1 ____ U1 ____ L2 ____ U2 ____ L3 ____ U3 ____

PK - 1 Information

Abrupt Initiation (Power) 100

Abrupt Termination (Power) 0

Abrupt Boundary Discontinuities LB Step I UB Step I

Shapes I/RP Slopes I/0 Slopes I/0

PK - 2 Information

Abrupt Initiation (Power) ____

Abrupt Termination (Power) ____

Abrupt Boundary Discontinuities LB Step ____ UB Step ____

Shapes ____ Slopes ____ Slopes ____

PK - 3 Information

Abrupt Initiation (Power) ____

Abrupt Termination (Power) ____

Abrupt Boundary Discontinuities LB Step ____ UB Step ____

Shapes ____ Slopes ____ Slopes ____

PM_i 6 PM_i SEQ NO 1

PM_{i+1} 2 SAMPLE WORD 5-UR

Duration Ratio 1/6

Power Ratio 25/42

Power Distribution 100 ___ ___ / 12 88 ___

Merge; Split 0 0 Boundary Revisions (Including Those Renamed in PM_{i+1})

DESTROYED	L1	___	U1	___	L2	___	U2	___	L3	___	U3	___
CREATED	L1	___	U1	___	L2	___	U2	___	L3	___	U3	___

PK - 1 Information

Abrupt Initiation (Power) 12

Abrupt Termination (Power) 0

Abrupt Boundary Discontinuities LB Step I UB Step I

Shapes I/1D Slopes I/0 Slopes I/0

PK - 2 Information

Abrupt Initiation (Power) 0

Abrupt Termination (Power) 0

Abrupt Boundary Discontinuities LB Step 0 UB Step +3

Shapes I/C Slopes 0/+.33 Slopes 0/+.17

PK - 3 Information

Abrupt Initiation (Power) ___

Abrupt Termination (Power) ___

Abrupt Boundary Discontinuities LB Step ___ UB Step ___

Shapes ___ Slopes ___ Slopes ___

PM_i 2 PM_i SEQ NO 2

PM_{i+1} 2 SAMPLE WORD 5-WR

Duration Ratio 6/5

Power Ratio 43/43

Power Distribution 12 88 / 12 88

Merge; Split 0 0 Boundary Revisions (Including Those Renamed in PM_{i+1})

DESTROYED	L1		U1		L2		U2		L3		U3	
CREATED	L1		U1		L2		U2		L3		U3	

PK - 1 Information

Abrupt Initiation (Power) 0

Abrupt Termination (Power) 0

Abrupt Boundary Discontinuities LB Step 0 UB Step 0

Shapes 1D/1D Slopes 0/0 Slopes 0/0

PK - 2 Information

Abrupt Initiation (Power) 0

Abrupt Termination (Power) 0

Abrupt Boundary Discontinuities LB Step 0 UB Step 0

Shapes 1D/1D Slopes +.33/-.20 Slopes +.17/-.20

PK - 3 Information

Abrupt Initiation (Power)

Abrupt Termination (Power)

Abrupt Boundary Discontinuities LB Step UB Step

Shapes Slopes Slopes

PM_i 2 PM_i SEQ NO 3

PM_{i+1} 4 SAMPLE WORD 5-WR

Duration Ratio 5/14

Power Ratio 41/34

Power Distribution 12 88 / 100

Merge; Split 0 0 Boundary Revisions (Including Those Renamed in PM_{i+1})

DESTROYED L1 ____ U1 ____ L2 ____ U2 ____ L3 ____ U3 ____

CREATED L1 ____ U1 ____ L2 ____ U2 ____ L3 ____ U3 ____

PK - 1 Information

Abrupt Initiation (Power) 0

Abrupt Termination (Power) 12

Abrupt Boundary Discontinuities LB Step I UB Step I

Shapes 1D/I Slopes 0/I Slopes 0/I

PK - 2 Information

Abrupt Initiation (Power) 0

Abrupt Termination (Power) 0

Abrupt Boundary Discontinuities LB Step -1 UB Step +4

Shapes 1D/BP Slopes -.2/0 Slopes -.2/0

PK - 3 Information

Abrupt Initiation (Power) ____

Abrupt Termination (Power) ____

Abrupt Boundary Discontinuities LB Step ____ UE Step ____

Shapes ____ Slopes ____ Slopes ____

PM_i 4 — PM_i SEQ NO 4

PM_{i+1} 3 — SAMPLE WORD S-UR

Duration Ratio 14/16

Power Ratio 34/30

Power Distribution 100 ___ ___ / 50 33 17

Merge; Split 0 0 — Boundary Revisions (Including Those Renamed in PM_{i+1})

DESTROYED	L1	___	U1	___	L2	___	U2	___	L3	___	U3	___
CREATED	L1	___	U1	___	L2	___	U2	___	L3	___	U3	___

PK - 1 Information

Abrupt Initiation (Power) 50

Abrupt Termination (Power) 0

Abrupt Boundary Discontinuities — LB Step I — UB Step I

Shapes I/10 — Slopes I/0 — Slopes I/0

PK - 2 Information

Abrupt Initiation (Power) 0

Abrupt Termination (Power) 0

Abrupt Boundary Discontinuities — LB Step +3 — UB Step I

Shapes I/SP — Slopes 0/0 — Slopes I/0

PK - 3 Information

Abrupt Initiation (Power) 0

Abrupt Termination (Power) 0

Abrupt Boundary Discontinuities — LB Step I — UB Step -5

Shapes I/SP — Slopes I/-.06 — Slopes 0/-.06

PM_i 3 PM_i SEQ NO 5

PM_{i+1} 3 SAMPLE WORD 5-WR

Duration Ratio 16/9

Power Ratio 29/29

Power Distribution 52 34 14 / 52 34 14

Merge; Split 0 0 Boundary Revisions (Including Those Renamed in PM_{i+1})

DESTROYED	L1		U1		L2		U2		L3		U3	
CREATED	L1		U1		L2		U2		L3		U3	

PK - 1 Information

Abrupt Initiation (Power) 0

Abrupt Termination (Power) 0

Abrupt Boundary Discontinuities LB Step 0 UB Step 0

Shapes 1D/1D Slopes 0/0 Slopes 0/+.33

PK - 2 Information

Abrupt Initiation (Power) 0

Abrupt Termination (Power) 0

Abrupt Boundary Discontinuities LB Step 0 UB Step 0

Shapes SP/SP Slopes 0/-.55 Slopes 0/-.55

PK - 3 Information

Abrupt Initiation (Power) 0

Abrupt Termination (Power) 0

Abrupt Boundary Discontinuities LB Step 0 UB Step 0

Shapes BP/BP Slopes -.06/0 Slopes -.06/0

PM_i 3 PM_i SEQ NO 6

PM_{i+1} 3 SAMPLE WORD S-UR

Duration Ratio 9/5

Power Ratio 39/31

Power Distribution 59 31 10 / 35 52 13

Merge; Split 1 1 Boundary Revisions (Including Those Renamed in PM_{i+1})

DESTROYED L1 0 U1 1 L2 1 U2 0 L3 0 U3 0

CREATED L1 0 U1 1 L2 1 U2 0 L3 0 U3 0

PK - 1 Information

Abrupt Initiation (Power) 0

Abrupt Termination (Power) 0

Abrupt Boundary Discontinuities LB Step 0 UB Step I

Shapes I/BP Slopes 0/0 Slopes I/-.2

PK - 2 Information

Abrupt Initiation (Power) 0

Abrupt Termination (Power) 0

Abrupt Boundary Discontinuities LB Step I UB Step 0

Shapes I/BP Slopes I/+.4 Slopes -.55/0

PK - 3 Information

Abrupt Initiation (Power) 0

Abrupt Termination (Power) 0

Abrupt Boundary Discontinuities LB Step 0 UB Step 0

Shapes BP/BP Slopes 0/-.2 Slopes 0/-.4

PM_i 3 PM_i SEQ NO 7

PM_{i+1} 4 SAMPLE WORD S-WR

Duration Ratio 5/5

Power Ratio 24/13

Power Distribution 37 50 13 / 100

Merge; Split 0 0 Boundary Revisions (Including Those Renamed in PM_{i+1})

DESTROYED L1 ____ U1 ____ L2 ____ U2 ____ L3 ____ U3 ____

CREATED L1 ____ U1 ____ L2 ____ U2 ____ L3 ____ U3 ____

PK - 1 Information

Abrupt Initiation (Power) 0

Abrupt Termination (Power) 37

Abrupt Boundary Discontinuities LB Step I UB Step I

Shapes BP/I Slopes 0/I Slopes -.2/I

PK - 2 Information

Abrupt Initiation (Power) 0

Abrupt Termination (Power) 0

Abrupt Boundary Discontinuities LB Step 0 UB Step +5

Shapes BP/BP Slopes +.4/0 Slopes 0/0

PK - 3 Information

Abrupt Initiation (Power) 0

Abrupt Termination (Power) 13

Abrupt Boundary Discontinuities LB Step I UB Step I

Shapes BP/I Slopes -.2/I Slopes -.4/I

PM_i 4 PM_i SEQ NO 8

PM_{i+1} 5 SAMPLE WORD S-WR

Duration Ratio 5/5

Power Ratio 13/5

Power Distribution 100 ____ ____ / 5 ____ ____

Merge; Split 0 0 Boundary Revisions (Including Those Renamed in PM_{i+1})

DESTROYED	L1	____	U1	____	L2	____	U2	____	L3	____	U3	____
CREATED	L1	____	U1	____	L2	____	U2	____	L3	____	U3	____

PK - 1 Information

Abrupt Initiation (Power) 0

Abrupt Termination (Power) 100

Abrupt Boundary Discontinuities LB Step I UB Step I

Shapes BP/I Slopes 0/I Slopes 0/I

PK - 2 Information

Abrupt Initiation (Power) ____

Abrupt Termination (Power) ____

Abrupt Boundary Discontinuities LB Step ____ UB Step ____

Shapes ____ Slopes ____ Slopes ____

PK - 3 Information

Abrupt Initiation (Power) ____

Abrupt Termination (Power) ____

Abrupt Boundary Discontinuities LB Step ____ UB Step ____

Shapes ____ Slopes ____ Slopes ____

APPENDIX C - NUMERICAL EXAMPLE OF DISCRIMINANT SET FORMATION

This appendix presents a numerical example of the algorithms defined on pages 77-83. For the sake of clarity, hypothetical RO's with only 12 elements are used. Each ro_j element may have the values 0-10, I (indeterminate), or blk (blank). The data is meant to represent the comparable RO's from five speakers' (m_1, m_2, m_3, m_4, m_5) single utterances of two different words (Word A, Word B). The transitions between PM_i and PM_{i+1}, RO_i^A and RO_i^B are the comparisons across speakers for the same word and define the sets J_s^A, J_d^A, J_b^A, J_s^B, J_d^B, J_b^B. Comparison of the RO_i^A and RO_i^B produce the sets:

Type I J_s^{AB}

Type II J_{II}^A, J_{II}^B, $cf_{J_{II}^{AB}}$, $cs_{J_{II}^A}$, $cs_{J_{II}^B}$

Type III J_r^{AB}

The power to discriminate between A and B resides in the Type II sets.

Samples from Word A

$PM_i \; RO_i \; {}^{A}m_j \; PM_{i+1}$

Samples from Word B

$PM_i \; RO_i \; {}^{B}m_j \; PM_{i+1}$

		Speakers					Resulting		Resulting	Speakers					
		m_1	m_2	m_3	m_4	m_5	RO_i^A		RO_i^B	m_1	m_2	m_3	m_4	m_5	
J_s^A	ro_1	I	I	I	I	I	I		I	I	I	I	I	I	ro_1 J_s^B
	ro_2	7	8	6	7	8	7.2	J_s^{AB} Type I	7.4	6	7	8	8	8	ro_2 J_s^B
J_{II}^A	ro_3	2	3	2	4	2	2.6	$cf\, J_{II}^{AB}$ Type II	9.4	9	10	9	9	10	ro_3 J_{II}^B
	ro_4	8	9	10	8	9	8.8		2.8	3	3	2	4	2	ro_4 J_{II}^B
	ro_5	I	I	I	I	I	I	$cs\, J_{II}^A$ Type II	X	1	10	3	6	9	ro_5 J_d^B
J_d^A	ro_6	1	blk	7	I	4	X	$cs\, J_{II}^B$ Type II	1.4	1	2	1	1	2	ro_6 J_s^B J_{II}^B
	ro_7	10	2	I	3	7	X		5.2	5	6	4	6	5	ro_7 J_s^B J_{II}^B
	ro_8	2	5	1	blk	7	X		X	3	9	2	blk	7	ro_8 J_d^B
	ro_9	10	9	blk	1	4	X	J_r^{AB} Type III	X	10	4	1	I	3	ro_9 J_d^B
J_b^A	ro_{10}	blk	blk	blk	blk	blk	blk		Y	2	9	10	1	3	ro_{10}
J_d^A	ro_{11}	2	10	blk	3	7	X		blk	blk	blk	blk	blk	blk	ro_{11} J_d^B
J_b^A	ro_{12}	blk	blk	blk	blk	blk	blk	J_s^{AB} Type I	blk	blk	blk	blk	blk	blk	ro_{12} J_d^B

Tolerance constant $\gamma = 2$

Numerical Example of the Rules for Generating the Disciminant Sets