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.

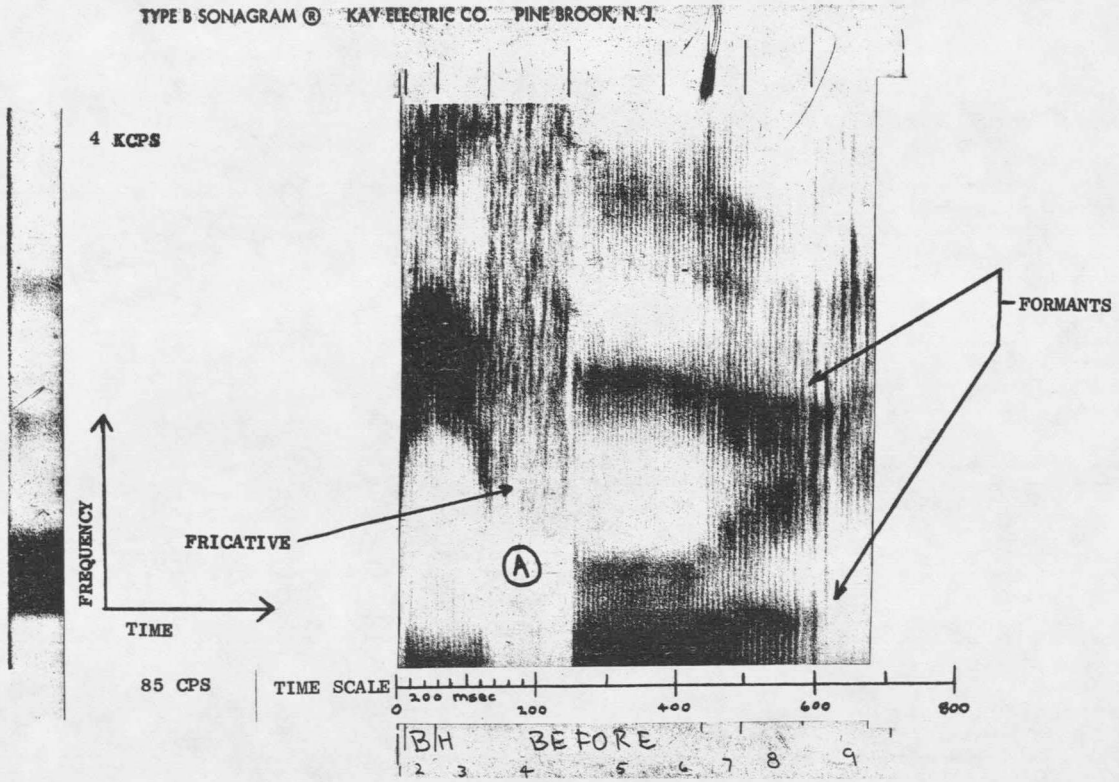

Figure 1. Sample Sonagram of /Before/.

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.