AD 437324

REPRODUCTION QUALITY NOTICE

This document is the best quality available. The copy furnished to DTIC contained pages that may have the following quality problems:

- Pages smaller or larger than normal.
- Pages with background color or light colored printing.
- Pages with small type or poor printing; and or
- Pages with continuous tone material or color photographs.

Due to various output media available these conditions may or may not cause poor legibility in the microfiche or hardcopy output you receive.

☑ If this block is checked, the copy furnished to DTIC contained pages with color printing, that when reproduced in Black and White, may change detail of the original copy.

AFCRL - 64 - 85

MULTIDIMENSIONAL MODEL FOR AUTOMATIC SPEECH RECOGNITION

B. V. BHIMANI

BHIMANI RESEARCH ASSOCIATES
1838 Massachusetts Avenue
Lexington 73, Massachusetts

FINAL REPORT

Contract No. AF 19(628)-2766
Project 4610
Task 461002

February 14, 1964

Prepared for:

AIR FORCE CAMBRIDGE RESEARCH LABORATORIES
OFFICE OF AEROSPACE RESEARCH
UNITED STATES AIR FORCE
BEDFORD, MASSACHUSETTS

Requests for additional copies by Agencies of the Department of Defense, their contractors, and other Government agencies should be directed to the:

DEFENSE DOCUMENTATION CENTER (DDC)
CAMERON STATION
ALEXANDRIA, VIRGINIA

Department of Defense contractors must be established for DDC services or have their 'need-to-know' certified by the cognizant military agency of their project or contract.

All other persons and organizations should apply to the:

U. S. DEPARTMENT OF COMMERCE
OFFICE OF TECHNICAL SERVICES
WASHINGTON 25, D. C.

AFCRL - 64 - 85

MULTIDIMENSIONAL MODEL FOR AUTOMATIC SPEECH RECOGNITION

B. V. BHIMANI

BHIMANI RESEARCH ASSOCIATES
1838 Massachusetts Avenue
Lexington 73, Massachusetts

FINAL REPORT

Contract No. AF 19(628)-2766
Project 4610
Task 461002

February 14, 1964

Prepared for:

AIR FORCE CAMBRIDGE RESEARCH LABORATORIES
OFFICE OF AEROSPACE RESEARCH
UNITED STATES AIR FORCE
BEDFORD, MASSACHUSETTS

ABSTRACT

The purpose of this study is to provide a theoretical basis for a general purpose speech recognizer. The research has focused upon the nature of normal speech, which can be distinguished from discrete articulation by the continuous movement (in normal speech) of articulators from one position to another; as a result, sounds in continuous speech are more likely to modify the production of surrounding sounds than they are in discrete speech.

Assuming that, according to the ergodic theory, sound changes occurring in everyday speech reflect and repeat the changes which have occurred in the historical development of language (because the physical modes of speech production are the same), linguistic examples and theories of sound change were studied. From this study, a body of rules for sound change or euphonic combination was derived and their applicability to the English language tested. These rules represent an error-correcting code to restore omitted or indefinite word boundaries and/or to restore the orthographic phone classes which are altered in continuous speech.

The study required the evaluation of existing research and theories, as well as the generation of some original data, the latter consisting of high-quality recordings of continuous speech samples. Both original data and previously published data were subjected to acoustic analysis of minute portions of the speech waveform. These measurements both suggested and justified a principle of segmenting speech, to be used in conjunction with the representations of speech sounds in the multidimensional model (according to the degree of freedom in various dimensions of their production), and the above mentioned error correcting code, to delineate a new conception of a general purpose recognizer.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

PREFACE

In order to provide a theoretical basis for a general purpose recognizer, we have investigated the possibility of organizing the information available on the various aspects of speech into a multidimensional model, based upon genetive linguistic, phonetic, and acoustic considerations. We have thus attempted to establish an orderly method for representing speech sounds. This orderly method is unique, we believe, for while it can readily be used to describe the sounds that occur in carefully and discretely articulated speech, it can also provide a basis for a recognition program for imperfectly articulated continuous speech. For example, the recognition of bet you (bechyou) in continuous speech utilized information which is similar to that previously known: the representation of be and the representation of chew. It is our rules of euphonic combination which bridge the gap between what was previously known about discrete speech (be and chew) and what we have discovered about continuous speech (bechyou), by indicating that such a modification of discrete speech is likely to occur in continuous speech.

Instead of undertaking the formidable task of examining vast samplings of continuous speech, we have constructed our model on the basis of existing literature. The physical basis of articulation has been and is currently being investigated thoroughly by other researchers. For the most part, our explanations of the physical production of sounds concur with widely-accepted descriptions: our one exception (and thus our major contribution) to this description is our emphasis on the distinctive nature of continuous speech. Existing theories suggest that normal speech can be reduced to its scientific essentials by studying the production of individual sounds, and then combining sounds in an additive fashion. That is, by forcing air through the articulators, sound is produced; changing the position of the articulators changes the acoustic properties of the sound. Thus for each arrangement of the articulators by a particular person, there corresponds a sound of reasonably distinct acoustic properties.

We contend, however, that one cannot merely use the sum of a sequence of separately - produced sounds to describe what happens in normal or continuous speech. For speech does not consist merely of placing the articulators in a position which is fixed for a particular sound, and forcing air through them, one breath for each sound. Instead, the air is forced through continuously, and the articulators are constantly moving from one position to another, making a continuous flow of sounds. It is then important to recognize that in continuous speech,

sounds can easily modify surrounding sound, so that the waveform of a sound produced in continuous speech can differ significantly from the waveform of that sound pronounced in isolation. In continuous speech, sounds can be eliminated, added, added together or substituted for one another.

Thus the matter of articulation, when applied to continuous speech is intimately connected with the phenomenon of sound change. It is on this basis that we undertook an historical survey of sound change in various Indo-European languages, which utilize the same physical modes of production. This study is presented in Sections 1 - 3 of this report. From this body of linguistic research, we collected several hundred tentative "rules" of sound change, assuming, by analogy with the Ergodic theory of physics, that all the sound changes which have occurred in the historical development of languages are being duplicated today, at a particular moment in a given language. We do not, however, accept such "rules" as final, until their occurrence in modernday English has been substantiated by examining samples of continuous speech.

Phonetic analysis is a tool of the linguist, and can be used only to ascertain how continuous speech is perceived by the human ear (i.e. whether or not words actually do or do not contain the sounds indicated in their orthography). As such, however, it provides a worthwhile indicator of the acoustic discrepancies which may occur in continuous speech.

The final analysis and criterion must be acoustic, however, for it is the acoustic waveform which must be understood by a speech recognizer. For this reason, Section 4 of this report, which presents acoustic evidence to justify our treatment of speech information, might be considered an essential contribution of this study.

Our work in this study has been limited to examining the characteristics of those sounds usually classified as consonants. (We do, however, make several general observations and recommendations about vowel treatment, although the vowels were not studied in depth.) In seeking to provide an orderly means of representing consonants, we have reached several major conclusions: (1) the character of a given consonant - its place or manner of articulation, and thus its acoustic representation - changes according to the sound which precedes or follows it. (2) For this reason, so-called "consonant clusters" should be treated as unique entities, not as the addition of two or more fixed sounds. (This is amplified in our discussion on segmentation in Section 4.) (3) If consonant clusters are treated as special consonants, then speech can be

divided into segments consisting of "consonant-vowel" combinations, including the onglide and offglide transitions to make recognition more precise.

The multidimensional model for speech recognition is thus an ordered manner of representing the various classes of consonants in such a way that a shift or drift of consonants to another class can be accounted for. The body of rules of sound change or Euphonic Combination, as we shall show, can be represented in symbolic form suitable for computer programming. Our model plus the rules of euphonic combination thus represents an error-correcting code for speech recognition. For since the same degrees of freedom exist - within the realm of physical possibility and necessity - we can predict the mistakes which may occur. We are thus operating by analogy with the Ergodic theory of physics, rather than following the hypothetical constructs of linguistics, which are at times contradictory and often unorganized.

Furthermore, our work has suggested segments which are better suited for recognition by the perceiver than either phonemes or words. And finally, our work has indicated the importance of including such aspects of speech as intensity and duration as considerations necessary for the segmentation of speech. Additional research in these areas seems advisable.

It may be noticed that certain of the concepts presented in this report will be familiar to the reader. We include such information for several reasons:

1) to state a common background and to provide information for those readers not specializing in any one of these aspects.

2) to provide detailed descriptions of our assumptions and thus to indicate the extent of applicability of our method and our results.

3) to order information which has been previously available from various sources, but which has never been presented in an organized form in the published literature.

4) to describe and to explain our method of ordering sounds, and to justify our positioning of sounds in the multidimensional model.

5) to ascertain that the just critics can find constructive aspects and that professional critics can be credited with justifiable comments.

The individual treatment of the work done by western linguists, of sound change, of the concept of the multidimensional model, and of the acoustic evidence substantiating this concept has necessitated a certain amount of repetition; such repetition is necessary for the sake of clarity. In order to organize and consider all the necessary aspects of speech, we have been denied a study in depth of several areas where such study seems advisable. Our thoroughness has been to include all aspects, rather than to examine certain of these aspects in great detail. Furthermore, a historic review was necessary for several reasons:

(1) To point out difficulties of definition which have confused previous research in this field. One serious example of unclear definition is the historic use of the term "phoneme."

(2) To place our work in perspective with contemporaries, and to clarify the stand taken in other published work.

(3) To evaluate which concepts in the published literature were irrelevant, and to determine which of these concepts could be adapted to suit our present needs.

Finally, qualifications of certain other aspects of our research must be touched upon.

While place and manner of articulation are used in defining phones, the acoustic waveforms are not said to depend on these aspects alone; this is recognized by our CV ordering of related aspects. The principal reason for their consistent use is the need for ordering information about phone combination which is available in linguistic literature where such nomenclature originated.

Our study introduces the importance of prosodic features of speech, such as duration and intensity, which have been too often ignored in work on automatic speech recognition, and which are not ever considered distinctive features when they actually do provide new differentia, as in balm and bomb.

The role of pitch, intensity, time normalization, etc., is left in a theoretical state primarily because of the need for depth studies in each of these areas for additional discussion. Moreover, the available information defines the situation only to the degree of justifying their inclusion as dimensions in the model.

It is worth noting, finally, that the results of a recent study, "Dimensions of Perception of Consonants" was published by Robert W. Peters in the December, 1963 Journal of the Acoustical Society of America (35:12, pp. 1985 - 9). In analyzing the psychological "distance"

between consonants, it was reported that certain dimensions could be ordered according to their importance in identifying consonants: manner of articulation was judged the most important dimension, followed by voicing, and then place of articulation, in that order.

We agree that these dimensions are important and necessary to the identification of consonants; indeed, these dimensions are included as the major criteria for categorization in each plane of our multidimensional model.

Section 1 of this report reviews the linguistic problems involved in a study such as this; Section 2 outlines the form of the Multidimensional Model; Section 3 considers sound changes derived from our linguistic phonetic and genitive research in terms of the model. It is in Section 3 (and Appendix H) that we propose a body of rules for euphonic combination in continuous speech.

Section 4 presents the acoustic data - the measurements made on portions of the speech waveform in order to test the validity of our approach.

Section 5 suggests a method of segmenting continuous speech into units which, when used in combination with the stored rules for euphonic combination, are suitable for computer processing. In that section we also perform a cursory examination of the frequency of occurrence for various rules of euphonic combination, and explore various techniques of computer formatting which could be used to match continuous speech to orthographic script.

MULTIDIMENSIONAL MODEL FOR AUTOMATIC SPEECH RECOGNITION

BY

B. V. BHIMANI

SECTION 1

LINGUISTIC ASPECTS OF SPEECH

GENERAL PROBLEMS OF AUTOMATIC SPEECH RECOGNITION

The basic purpose of this study is to define an orderly set of relationships that exist between speech patterns as they are physically produced and the sound of speech as it is perceived by human and mechanical means. It is our thesis that the sounds of human speech are not random phenomena, arbitrarily measured. Rather, there is a direct connection between the way speech is produced within the physical limits of choice available, and the sounds that the speaker produces. As a theoretical and practical aid in analyzing the related data of phonemic combination and acoustical perception we introduce the concept of a multi-dimensional model organized according to the physical freedoms a speaker may exercise in articulating his sounds and their interaction with each other.

Our study, it should be emphasized, represents both a synthesis of past works in phonetics, linguistics, and acoustics and the first general outline of what has been studied, what is relevant, and what is needed in the broad field of speech recognition. The work of contemporaries in investigating phenomena of acoustics of speech is considered as it relates to our concepts.

In combining such data we do not assert that speech can be defined and measured according to rigid rules of phonetic combination. Rather our problem is the relative freedom an individual speaker has to vary the sound of his words and still make them recognizable to the human ear. A model for speech recognition must have categories comprehensive enough to include these variations. It is particularly for this reason that we must study the physical causes of such variations, and measure sounds in units which will allow the greatest possible freedom in identifying related and unrelated phones.

As an aid to understanding this approach we include first a review of the history of phonetics and phonemics. A general background for our concepts is provided also by the charts illustrating the interaction of several of our dimensions, the general discussion of the divisions which our model makes and the reasons for making them; and the specific analyses of the problems involved in classifying and measuring each of the dimensions - manner of articulation, place of articulation, resonance, vowels, duration, intensity, and frequency.

In this discussion we have constantly related our formulation of how speech is produced to the ways of measuring speech perception, ranging from the human ear to spectral analyses and time-amplitude waveforms. Thus adequately outlined our formulation provides the basis for more detailed analysis of the problems involved in both its theoretical concepts and in its possible application in a practical field such as the mechanical development of a general purpose speech transcriber.

One of the basic problems in building an automatic speech recognizer is to decide what units the machine should recognize. We can choose between larger units, such as words and smaller units, such as sounds. When we consider the fact that the English language has several hundred thousand words, it seems impractical to build a word-recognizer. Once we have made the decision to build a sound-recognizer, we must decide how it will recognize sounds. It has been suggested that an automatic speech recognizer should be a phoneme recognizer. We do not agree with this suggestion, but before we can give our reasons, we must first discuss the meaning of "phoneme."

The word "phoneme" is currently used with at least two different meanings. Sometimes it is a synomym for "speech sound" and sometimes it refers to a class of speech sounds. It is primarily linguists who use the word in the latter sense, and since they do not always explain the term, it is frequently difficult for those who have not read widely in this field to follow the fine points of their discussions.

In this section we will describe the linguist's use of the word "phoneme" and the concepts which underlie it. The discussion will begin with a brief history of how the concepts evolved and a description of some currently-held theories about phonemes. We will next describe the techniques of phonemic analysis. Finally we will discuss the relevance of the phoneme to automatic speech recognition and explain why we think some other sound unit should be used for the machine.

I. HISTORY AND DESCRIPTION OF PHONEMIC THEORY

Linguistics as an academic discipline began in the early part of the nineteenth century with the discovery that there were regular sound correspondences between the Germanic languages, such as English and other members of the Indo-European group, such as Sanskrit, Greek, and Latin. These correspondences were of the type, Latin p corresponds to English f, Latin t corresponds to English th, Latin c (pronounced k) corresponds to English h. Some words illustrating these correspondences are:

Latin pater	English father
Latin tu	English thou
Latin tres	English three
Latin centum	English hundred

The linguists who discovered these correspondences explained them with the theory that most of the languages of Europe and many of the languages of Asia are descended from one single language which was spoken at some time in prehistory. This language was given the name Proto-Indo-European. Since it was spoken in prehistory all our knowledge of it comes from comparing the languages descended from it.

In comparing these languages to decide whether the Proto-Indo-European word for three began with t or th, the linguists concluded that the t is original and the th an innovation, because only the Germanic languages have th. (At present, only two of the Germanic languages have th, but we know from written records that the others had it earlier.)

The discovery that these sound changes had taken place was important not only for an understanding of language relationships, but also for an understanding of phonetics. p, t, and k are phonetically similar; they are all voiceless stops. f, th, and h are also phonetically similar; they are all voiceless continuants. This means that the process whereby p became f was identical with the process whereby t became th and k became h. This suggests that the sounds of a language operate as a system, rather than independently of each other.

There are two other important sets of sound-changes between Proto-Indo-European and early Germanic. They are illustrated by the following words:

Latin duo	English two
Latin id	English it
Sanskrit dha	English do

These correspondences are summarized by the statements that the Proto-Indo-European voiced unaspirated stops d, b, and g became the Germanic voiceless stops p, t, and k and that the Proto-Indo-European voiced aspirated stops bh, dh, and gh became the Germanic voiced stops b, d, and g or aspirants β, ð, and ɣ. (The exact nature of these sounds is not clear because the Germanic languages have made various sound changes since then. It is true, however, that Sanskrit dh usually corresponds to modern English d). Again, similar sounds underwent similar changes.

The discovery of these sound-correspondences gave the impetus for further research into the sound-correspondences among the Indo-European languages. All of the languages descended from Proto-Indo-European had made some sound changes, and the scholars' problem was to reconstruct the original language.

In comparing words to discover sound-correspondences, the scholars never knowingly compared words which one language had borrowed from another. Whenever such words were included, the results disagreed with the correspondences discovered by other comparisons. There are many words besides pater and father which show that Latin p corresponds to English f, but the English word paternal appears to show that Latin p corresponds to English p. The explanation is that paternal is borrowed from Latin, and therefore should not be used to discover the sound correspondence between English and Latin. The task of establishing the correspondences was complicated by the presence of loan-words which were not recognized as such, and by the fact that some languages had undergone many sound changes, and some of the later changes obscured the effects of the earlier ones.

There were three types of sound-change which these early nineteenth century linguists recognized; conditioned, unconditioned, and sporadic. A sound change which had taken place only under certain circumstances was a conditioned change. Proto-Indo-European t became English th everywhere except where another spirant or sibilant preceded it. Thus the t in Latin tu corresponds to the th in English thou, but the t in Latin sto corresponds to the t in English stand. A sound change which takes place under any circumstances is an unconditioned sound change. Proto-Indo-European o became a in the Germanic languages in all positions. Latin toga goes back to the same Proto-Indo-European word as English thatch; the word probably had the original meaning of a covering. The third type of sound change which the early nineteenth century linguists considered important was sporadic sound change. As the name implies, this type of sound change was described as not subject to any rules.

In 1876, the idea of sporadic sound change was attacked by a group of scholars who maintained that all sound change was regular, that is to say, all sound changes could be divided into two groups, those which were unconditioned and those for which the conditions could be clearly stated, if all necessary data were available. This theory was an innovation because up to that time the most widespread view about sound change was that each word had its own history; the new theory suggested that each sound had its own history. This theory not only had immediate applications to the problems of reconstructing Proto-Indo-European but it also had a long-range effect on all later theories about how sounds function in language.

At about the same time that the theory of regular sound change was finding acceptance, linguists also began to take interest in the fact that the

sounds used in human speech showed much more variety than anyone had given them credit for. The impetus for this came from the dialect geographers, who were making maps of regions or even whole countries which showed just how the speech of the people in each village differed from that of the people in the adjacent village. Once they became aware of the great phonetic variety, the linguists began making phonetic transcriptions which faithfully recorded every variation they could hear. This approach became the standard practice, and since it is true that no two speech events by the same speaker are absolutely identical and many of the differences are great enough to be audible, it became common to indicate in the transcription that a given individual's utterance of the word "cat" on Monday differed slightly from his utterance of the same word on Wednesday.

There was another reason for variations in phonetic transcriptions, although most of the phoneticians of that time did not realize it. No human being, however carefully trained, can function as an automatic transcribing machine. The listener shows variations just as the speaker does. On one day a sound might seem to be the vowel of bet somewhat lowered, and on the next day it might seem to be the vowel of bat somewhat raised. The situation is not improved by training the linguist to hear seven different steps along the continuum from high vowel to low. Instead of transcribing a sound as a low [ɛ] one day and a high [e] another day, the linguist transcribes it as a vowel of the fourth highes step on one day and the fifth highest on another. It is true that some linguists are almost perfectly consistent in their transcriptions, but this consistency appears to be an inborn gift rather than something which can be taught.

The extremely detailed transcriptions produced by attempting phonetic accuracy were not easy to work with. While it was true that the word "cat" was never pronounced exactly the same way twice, it was also true that it was always recognizable as "cat" and not confused with "bat" or "gat" or "at." It became clear that some phonetic differences played an important role in a given language while others were irrelevant. With this realization came the need for terminology which would make it easy to talk about the distinction between relevant and irrelevant differences. It was at this point that the word "phoneme" came into common use, to refer to a group of sounds, differences among which were irrelevent.

Although it has been widely used for about forty years, there is still no agreement on precisely how to define the term "phoneme." This is not to say that no definitions have been proposed; there have been many definitions and much discussion, but they have not let to unanimity.

The different definitions of the phoneme which have been proposed fall into four main groups. Those of the first say that the phoneme is a psychological entity; it is a physical entity to those of the second; and it is both a physical and a psychological entity to those of the third; whereas those of the fourth say that phoneme is a class of sounds, but they do not ascribe physical or psychological reality to this class. The linguists who subscribe to definitions of the fourth type do not flatly state that the phoneme has no physical or psychological reality; they say that there is not sufficient proof on this point, and until proof is forthcoming, it is better not to make unnecessary assumptions.

It should be noted that although there is disagreement on how to define the phoneme, it is possible and even common for linguists who define it differently to arrive at the same phonemic analysis of a given set of data. This is because the techniques of phonemic analysis are quite similar, no matter what it is that the linguist thinks he is analyzing.

This is not to say that all phonemic analyses of the same body of data will be the same; they will not be. However, the differences cannot be wholly attributed to differences in phonemic theory. Yuen-Ren Chao gives the following list of criteria which are used for phonemic analysis (Chao, 1934).

(1) phonetic accuracy, or smallness of range of phonemes

(2) simplicity or symmetry of phonetic pattern for the whole language

(3) parsimony in the total number of phonemes

(4) regard for the feeling of the native speaker

(5) regard for etymology

(6) mutual exclusiveness between phonemes

(7) symbolic reversibility (that is to say, given any phonemic symbol in a language, the range of sounds it represents is determined; given any sound in the language, its phonemic symbol is determined).

Chao points out that different linguists don't attach the same weight to these criteria. Some of the criteria such as (5) are especially important to those who define the phoneme as a psychological entity; some, such as (1), are especially important to those who define the phoneme as a physical entity. There are others, however, such as (6) which are not weighted by

one or another of the four definitions listed. This means that differences in analysis of the same data arise from different definitions of the phoneme and from different criteria of phonemic analysis.

II. TECHNIQUES OF PHONEMIC ANALYSIS

Despite the disagreement about defining the phoneme and about the criteria for phonemic analysis, most linguists use similar procedures when making a phonemic analysis, and they come up with similar results. The first step in phonemic analysis is to make a phonetic transcription, but there is a problem connected with this.

As the linguist listens carefully to a native speaker, there seems to be an enormous number of different sounds. Since he knows that it is highly unlikely that two speech sounds will be physically identical he is not surprised at the number of differences he hears but it poses a serious problem in transcribing and analyzing the utterances. The customary solution is for the linguist to note only those differences which he can analyze. This means that he uses the same symbol to transcribe two sounds which he knows are different. Since they are different, they cannot be called the same sound, but it is convenient to have some term to refer to all sounds which are transcribed with the same symbol. We shall use the term phone class. It should be noted that the phone classes of one linguist will not necessarily coincide with those of another. The number and extent of the phone classes of any linguist depend on his training, on his inborn ability to analyze sound differences, and differences that are relevant in his native language.

If a linguist transcribes very few phone classes because he can analyze few differences, he may have difficulty making a phonemic analysis. If he places two sounds with relevant (phonemic) differences in the same phone class, he is in trouble. It is essential for phonemic analysis that all sounds which are phonemically different shall have different phonetic transcriptions. It is for this reason that the linguist notes all the differences he can analyze; he does not know which are phonemic and which are not.

There are two types of non-phonemic variation. One type results from the fact that two speech sounds are almost never physically identical. The variations of this type are minor. The second type results from the fact that it is common in language for a phoneme to have quite different phonetic realizations in different environments; these are called allophones. In English the k of key is different from the k of coo, although they both belong to the /k/ phoneme. The k of key has its place of articulation at the front part of the /k/ range, because the following vowel is a front one. The k of coo has its place of articulation at the

back part of the /k/ range, because the following vowel is a back one. These two [k] 's belong to the same phoneme in English because they are phonetically similar and do not contrast. There are no minimal pairs identical except that one word has the [k] of key where the other has the [k] of coo. In Rumanian, however, these two [k] 's belong to different phonemes, and there is at least one minimal pair, identical in every respect except that one word has the front [k] where the other has the back [k] . The words are cu 'with' and chiu 'cry'. The letter c is used to spell the front [k] .

In Appendix A we analyze a small amount of phonetic data; this is to show how a linguist decides which differences are phonemic and which are non-phonemic. As Appendix A shows, the Vietnamese [k] and [p] alternate according to the adjacent vowel, just as the two [k] 's do in English. In another language which has these two sounds, they may belong to different phonemes. We cannot predict these matters from one language to another. The phoneme is a structural unit of language, and the phonetic shape of each phoneme must be determined for each language.

In this section we have described how sounds function in relation to language; we shall next consider briefly the question of how an automatic speech recognizer should function in relation to speech sounds. Any design for an automatic speech recognizer must, first, be feasible from the engineering point of view, and, second, must distinguish between all utterances which are different in the language being analyzed. It should be noted that the second condition states only that a speech recognizer should distinguish between all different utterances; we do not stipulate that it must ignore non-phoneme differences. We doubt that it is practical to build a phoneme recognizer.

Phonemic analyses are made by human beings with all the resourcefulness of the human brain. No one knows how the brain works, but we do know that it is the most efficient self-adjusting mechanism in the world. Speech takes advantage of this fact by requiring listeners to ignore some details and concentrate on others. The problem in automatic speech recognition is that we do not know how to build a machine that is equally self-adjusting. A human being, hearing a specific sound feature, can decide whether it is important and should be noted or it is irrelevant and should be ignored. In other words, he can decide, as each sound segment occurs, which aspects he should pay attention to and which he should ignore. He may pay attention to a feature in one context and ignore it in another. A machine could do this only if it were given a precise description of the circumstances under which the feature should be noted or ignored. If we give a machine such a description (assuming that such a description is possible), we render it incapable of making any adjustment whatever for the assimilations which occur when one sound follows another. If we do not tell it to ignore certain features under certain circumstances, then the number of different sounds

which it recognizes will be greater than the number of different sounds which a native speaker of the language needs to recognize. However, given the choice between a machine which is not flexible and one which makes some distinctions which are non-phonemic and perhaps "unnecessary", the authors at present believe that the machine which perceives non-phonemic distinctions is to be preferred.

In all languages, there are some phonetic variations which are the result of one sound influencing a nearby sound. Some of these occur every time, and thus they are predictable; others occur at some times and not at other times, and thus they are not predictable. The front [k] of key is predictable; it always occurs before front vowels. The voiceless dental stop [t̪] which sometimes replaces the English voiceless dental spirant [þ] (as in thin) is not predictable. The phrase with care is pronounced sometimes with [þ] at the end of with and sometimes with [t̪]. Since we cannot give a machine a precise rule about the substitution of [t̪] for [þ], the solution is to let the machine recognize [t̪] as a separate entity. When it fails to find [Wit̪] in its dictionary, it will check with the ordered set of sounds (described in the following sections), note that [t̪] is only one leaf removed from [þ], and decide that the work in question is with.

Thus, although in Vietnamese (see Appendix A for details) [k] and [p̚] are considered in at least one analysis to be phonemically the same, we will not require an automatic speech recognizer to recognize them as the same sound. We require that the machine distinguish between final [p] and [p̚] because the language contains the words [tɔp] and [tʌp̚] which must be distinguished. But this is our only stipulation about the recognition of final stop consonants in this language. We need not instruct the machine to note the p closures of [p] but ignore the p closure of [p̚].

We propose, in general, that the machine recognize sound segments as a linguist recognizes "phone classes" in analyzing an unknown language. These basic machine acoustic segments will be further described in Sections 4 and 5; for the present it is sufficient to emphasize that these segments will be designed to distinguish non-phonemic differences, such as ki and ko mentioned above.

We propose, then, that the machine recognize phone classes as the linguist does when he is analyzing a new language. These phone classes will not necessarily coincide with those that any linguist would use, any more than the phone classes of one linguist necessarily coincide with those of another. The machine will probably have more phone classes than most linguists. It will also have a complete dictionary of the language, written in its phonetic script. Whenever it receives a word which is not in its dictionary, it will refer back to its rules and to the matrix to determine

what sound-substitution has taken place.

There is an added advantage to having the machine recognize phone classes. We can switch it from one language to another by giving it another dictionary and set of rules. The phone classes it recognizes will remain the same. By designing a phone class recognizer we achieve a flexibility and reliability which, although they do not match those of the human brain, are considerably greater than those which could be achieved by a phoneme recognizer.

SECTION 2: GENERAL DISCUSSION OF THE MULTIDIMENSIONAL MODEL

INTRODUCTION

The purpose of the model is an orderly presentation and analysis of the phenomena of human speech. In effect we are trying to represent how the different freedoms which every speaker has in shaping the sounds of his words can be classified to allow for the possibilities of individual sound variations, then presented in a regular order so that they may easily be interpreted by human or mechanical means.

In articulating a sound the individual speaker has a number of freedoms in its production. On the other hand, he is limited to a given combination of influences he may use at any one time. By identifying the major sources of physical action which a speaker may use to produce a sound, then studying their influence upon each other, we come to an orderly formulation of how speech is produced.

In our model each dimension signifies an independent choice which a speaker makes in uttering a sound or a phrase. He has the freedom to choose where he will place his tongue in his mouth when uttering a sound (place of articulation); how he will move his tongue and lips to shape the sound (manner of articulation); and whether to use only his mouth as an echo chamber or whether to add the vocal flaps and the nasal passages (resonance); in pronouncing vowels; moreover, a speaker may decide how to move his cheek and jaw muscles, thereby determining which vowel he is enunciating.

In addition to these physical means of producing speech a speaker may use subtle variations of emphasis. They include how long to take in pronouncing a sound (duration); what sounds to stress (intensity); and where to vary pitch for emphasis (frequency); such freedoms qualify as dimensions particularly because they can act as independent agents in modifying the acoustic waveforms of speech. Moreover, they represent a special type of modification conveniently studied in a separate category.

Having provided a framework for analyzing speech patterns, we face two further tasks, discussed with each dimension. The first is to account for variations of choice that occur in normal speech -- it is a tautology for instance that an individual seldom pronounces a word identically twice in a row, nor do any two people pronounce the same word exactly alike. While furnishing classifications adaptable enough to cover such variations, however, we must always keep in sight our main objective -- that such a model be able to be developed for practical use.

To accomplish both ends the following discussion constantly relates the work of measuring speech phonetically to the equally important task of processing speech patterns by mechanical means.

I. DISCUSSION OF THE DIMENSIONS

Appendix B shows a chart showing the interaction of several separate dimensions of our model. The horizontal axis showed how a certain sound might be heard when pronounced at the same place in the mouth but given different resonances. The vertical axis showed the effect different positions of the tongue (place of articulation) would produce within the same resonance. Below the first chart we have indicated on a separate chart how use of the tongue in a given position can further modify a sound (manner of articulation).

All the sounds indicated on the front plane are sounds we define as consonants. (For the sake of clarity the consonants on the front plane are represented on separate leaves, which correspond to the consonant classes -- stop, nasal, sibilant, affricate, etc.) These sounds, however, are further modified by vowels that follow the consonants. An open a, for example, will modify almost every preceding consonant shown on the front chart. For this reason we have indicated on the z axis a series of planes showing the sounds of consonants as they are heard when articulated with given vowel sounds: k becomes ku, ka, ko; g becomes gu, ga, go, etc.

Only a few dimensions are shown on our charts, it should be noted, and these primarily as an illustration of how the separate dimensions can be related to each other for machine identification. Each of the following dimensions, however, represents an independent method by which the sounds of a voice may be varied; every such category must be studied separately and given special treatment in the construction of our model.

Having illustrated how the interaction of the separate dimensions can be represented, we may turn to a more specific examination of how the sounds thus graphed are physically produced and how they can then be turned into signals that a machine can classify. This we will do in our following discussion of the dimensions.

A. MANNER OF ARTICULATION, PLACE OF ARTICULATION, AND RESONANCE

Manner of articulation, place of articulation, and resonance are all physical modes through which men can alter their speech so that it may produce new information-bearing varieties of sound for the human

ear. Equally important, these modes of change need to be measured by mechanical means, thereby creating an opportunity for the transcribing machine to recognize the same sounds as the human ear.

In this section, accordingly, we must give equal attention first to the many physical ways of varying sounds for the human ear and then to the limited number of ways a machine may record them. For machines there are two principal ways that such sounds can be represented. The first is spectral analysis of energy concentrations in the speech wave. The second is time-amplitude plots of the changing shape of the sound-wave that occurs in a given word.

In the subsection below sounds are classified according to their similarity of physical production (as indicated in the X-ray photographs). They must then be measured according to their physical effect on the facilities of a transcribing machine (as indicated in the spectrograms and time-amplitude plots.)

The characteristic patterns of spectrographs and time-amplitude plots are particularly important in helping the machine identify manner of articulation: thus we include examples of both measurements as well as illustrative X-ray photographs for most phone classes in subsection A. Place of articulation, however, can be distinguished primarily by its formant transitions to and from adjacent vowels. These transitions are considered in our following treatment of vowels. For this reason we have included only a few illustrative figures for place of articulation and resonance.

It should finally be noted that the following three topics are separate dimensions. They are discussed together for convenience in relation to the preceding chart showing their specific interaction in helping our model translate physical combinations of sound into patterns identifiable by a machine. (See Appendix B).

1. Manner of Articulation

The distinguishing characteristic of this dimension is how the tongue and lips are used in producing a sound. The dimension has five subdivisions. These are presented in the following chart, with examples of the phones which fall into these subdivisions. For each example the usual English spelling, the phonetic symbol and a word containing the sound are given.

Manner of Articulation	Example		
	Usual Spelling	Phonetic Symbol	Word Containing the example
(1) Stops	p, d	p, d	pin, din
(2) Spirants	f, th*	f, ð	fin, the
(3) Nasals	n, ng*	n, ŋ	sin, sing
(4) Sibilants	z, sh*	z, ʃ	zoo, she
(5) Affricates	ch*, j	č, ǰ	chin, join
(6) Laterals	l, l	L, l	dull, let

* It should be noted that although two letters are needed to spell this sound in English, it is not two sounds, but one. This is what phoneticians call a digraph.

The characteristics of the subdivisions of this dimension are described below:

(1) Stops: The stops have a complete closure somewhere in the mouth for a brief period, which results in an almost complete cessation of sound. This cessation lasts about twenty milliseconds. The velum at the top of the throat closes during the articulation of a stop, so that no air escapes through the nose.

Gunnar Fant (Acoustic Theory of Speech Production, p. 186, Figures 2.6 - 8) presents X-rays of the side of the mouth: while the stop b is being pronounced, the point of contact is identical with that of the stop p, but the latter is not voiced.

A time-amplitude plot of the b sound (Figure 1) shows energy present even during closure. (This can also be detected in Ilse Lehiste and Gordon Peterson, "Studies of Syllable Nuclei 2", page 54, spectrogram of bit.); this is one way a machine might distinguish voiced stops from voiceless stops, because there is no energy present during closure in the case of voiceless stops.

(2) Spirants: The spirants are articulated by forming a constriction in the buccal or mouth cavity. In most cases this constriction is formed by raising some part of the tongue until it almost touches the roof of the mouth or the upper teeth. In the case of the labiodental spirants, however, such as f, or v, the constriction is formed by placing the upper teeth very close to the lower lip as in Gunnar Fant's X-ray tracing of f.

Figure 1

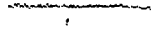


Figure 2

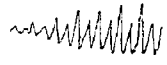
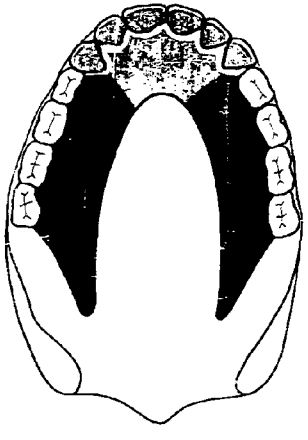
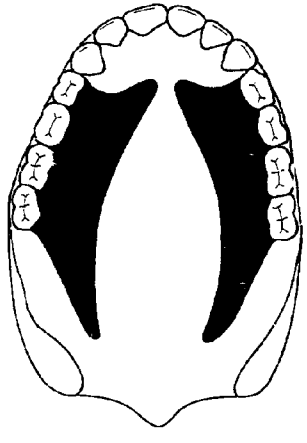


Figure 3. Time-Amplitude Waveform for m



TH



S
Z

Figure 6

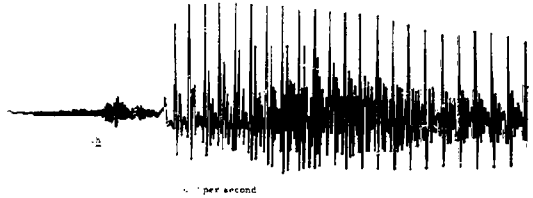


Figure 7

(Acoustic Theory of Speech Production, p. 170, Figure 2.6-1). The velum is closed for the articulation of a spirant. Both the spectrogram of v (Lehiste and Peterson, "Studies of Syllable Nuclei 2" p. 10, spectrogram of veal), and the time amplitude plot of f. (Figure 2) reveal a characteristic presence of random energy which a machine might identify.

(3) Nasals: Physically, nasals are articulated almost the same as stops; the only difference is that for a nasal the velum remains open, so that a stream of air escapes from the nose. Thus the articulation for m (Fant, p. 140, Figure 2.4-1), is the same as that for b (Fant, p. 186.) except for the velum at the back of the mouth. Accoustically, nasals differ from stops because there is no cessation of sound and no sharp burst of air when the closure is released.

These differences might be easily identified by a machine because nasalization creates an extra low-frequency formant on a spectrogram in addition to the three formants normally present in consonants and vowels (Lehiste and Peterson, p. 10, Spectrogram of "coin".) The time-amplitude plot, moreover, shows the relatively great intensity of low frequencies in nasal sounds, indicated by the comparative regularity and wide spacing of the peaks and valleys (Figure 3), although the peak radiated amplitude of nasals is significantly lower than that of many vowel sounds.

(4) Sibilants: The sibilants are like the spirants in that they involve a constriction in the mouth, but the shape of the tongue is different, and the characteristic sound is produced, not at the point of constriction, but when the air which has rushed through this constriction hits the upper front teeth (Von Essen, 1953, p. 73.)

As an illustration of this difference in pronunciation we include two palatograms of the spirant th (Figure 4) and the sibilants s or z (Figure 5). Black areas represent close contact between the tongue and the roof of the mouth, gray areas represent loose contact. The close contact and narrow opening of spirants as opposed to the loose contact and diffused opening of the sibilants is apparent. (For a description of how palatograms are made see Appendix C). Sibilants are acoustically different from spirants in that a sibilant has no energy below a certain frequency, but very high random energy above that frequency (as in the spectrogram of sh, (Lehiste and Peterson, p. 61, Spectrogram of shag) while spirants have much lower random energy spread throughout the spectrum. Time-amplitude plots of sh also reveal that sibilants have energy at higher frequencies than spirants do. (Figure 6). (According to Katherine S. Harris, 1953, sibilants also differ from spirants in that

they can be identified in listener tests without the help of transitional cues from the adjacent vowels, while spirants cannot be identified without transitional cues).

(5) Affricates: Physically affricates are produced by releasing the energy of a stop burst into the energy of a following sibilant. As the mouth is in almost constant motion we omit a physical diagram of its articulation; the reader can discover how affricates are formed by noting the similarity between white shoes and why choose.

In measuring affricates mechanically, traditional phonetics has assumed that the energy of the stop burst appears just at the beginning of the following sibilant. We have acoustic evidence, however, that the major release of the stop energy occurs not at the beginning of the sibilant, but in the middle of it. This can be seen in the time-amplitude plot (Figure 7). Information from the time-amplitude plot is more significant for our model than information from the spectrogram in characterizing affricates, because spectrograms are less able to record variations in sound-wave intensity.

(6) Laterals: In the articulation of a lateral, such as l, the middle of the tongue is in firm contact with the teeth or the roof of the mouth, and the air stream escapes at the side of the tongue. It has proved very difficult to give a clear acoustic description of l; researchers at Haskins Laboratories (O, Connor, Gerstman, Liberman, Delattre, and Cooper, 1957) also report difficulties in synthesizing it.

2. Place of Articulation

The place of articulation is that part of the buccal (mouth) cavity which has the smallest cross-section during the articulation of a specific sound. It is determined by the position of the articulators (the tongue or the lips) rather than how they are used in a given position. In this study we will consider six places of articulation; these are listed in the following chart with examples of the phones which fall into these subdivisions. The heading "Usual spelling" means usual English spelling.

Place of Articulation	Example		
	Usual Spelling	Phonetic Symbol	Word Containing the Example
(1) Guttural	k, c ("hard"), g ("hard")	k g	came game
(2) Palatal	sh	ʃ	shin
(3) Alveolar	t, d	t, d	tin, din
(4) Dental	th, th	θ, ð	*thin, *then
(5) Labiodental	f, v	f, v	fear, veer
(6) Labial	p, m	p, m	peer, mere

* It should be noted that thin and then do not have the same initial sounds although they start with the same letters; also both words have only one consonant before the vowel, although two letters are needed to spell the consonant.

(1) The Guttural Phones : Gutteral phones (i. e. speech sounds) have the smallest cross section in the back part of the mouth, as in the X-ray of k (Fant, p. 186, Figure 2.8-6). The constriction is achieved by raising the back part of the tongue.

(2) The Palatal Phones : The palatals have the smallest cross-section in the mid-part of the buccal cavity. This is achieved by raising the middle part of the tongue towards the hard palate (the middle part of the roof of the mouth).

(3) The Alveolar Phones: The alveolar phones have the narrowest cross-section at the alveolar ridge, just in back of the upper front teeth. The constrictiction is achieved by the bringing the tip of the tongue to the alveolar ridge.

(4) The Dental Phones: The dental phones have the narrowest cross-section at the upper front teeth. There are actually two places of articulation involved here: the tip of the tongue may be brought either to the back of the teeth, as in the X-ray of t (Fant, p. 186, Figure 2.8-6) or to their biting edges.

(5) The Labiodental Phones : The labiodental phones have the greatest constriction between the upper teeth and the lower lip: see the illustrations for [f] under the Spirants.

(6) The Labial Phones : The labial phones have the greatest constriction between the two lips: see the illustrations for [p] under the stops.

(7) Variability of Place of Articulation: Some English phonemes show much more variation in the place of articulation than others do. For most speakers, the place of articulation of all allophones of /p/ is approximately the same, but there are three places of articulation for the allophones of /l/ in leave, tilt, and milk, and the differences in pronunciation are audible; /p/ has a narrow range because there are two other voiceless stop phonemes in English, and the existence of these others limits the variability of /p/. Since /l/ is the only lateral phoneme in English, the sole limits on its variability are physiological; all parts of the tongue do not lend themselves equally well to lateral articulation. The extreme variability of /l/ probably explains why

researchers have not succeeded in discovering its acoustic characteristics.

The sibilants are another group of sounds which show variation in place of articulation, but the situation here is somewhat different. The sibilants are the only phone classes whose characteristic sounds do not originate at the point of constriction. We have already discussed the fact that the hissing sound of the sibilants arises when air which has been channeled through a narrow groove hits the teeth. The narrow groove is formed by the tongue, and the difference between [s] and [ʃ] (as in see and she) lies in the size of the groove: [s] has a smaller groove which is usually made with the tip of the tongue, and [ʃ] is made with the part of the tongue just behind the tip. The difference between [s] and [ʃ] can be analyzed either as a difference in manner ([ʃ] is articulated with a wider groove), or as a difference in place ([ʃ] is articulated further back). At present we choose to consider it a difference in place, but we may change our analysis later.

When we come to the rules of euphonic combination, the different degrees of variability will prove important.

3. Resonances and Aspiration

There are three types of resonance -- voiceless (only the friction of breath within the mouth); voiced (using vibration of the vocal flaps for modulating air that flows through the buccal cavity) and voiced plus nasal (coupling the nasal cavities with the previous system). In this subsection we also include aspiration because it occurs in English only with voiceless stops. As these resonances modify almost every sound that the mouth can physically produce, their identification in a separate dimension is important to our model.

(1) Aspiration. An aspirated phone is one which has an audible rush of air after the phone itself is articulated. It should be noted that the burst of sound when a stop is released is not aspiration; it is a necessary part of the articulation of the stop.

As indicated above, the contrast between aspirated and unaspirated sound is included with the resonances because in English only certain stops are aspirated. Stops which are voiceless (such as [p] are also aspirated; stops which are voiced (such as [b]) are unaspirated. The only exception to this is that [p] , [t] , and [k] , are not aspirated after [s]. This means that the p in pin is not like the p in spin.

There is one aspect of aspiration which requires particular study. The terms "aspirated" and "unaspirated" are usually applied only to stops, but the Sanskrit grammarians reported that their language had aspirated and unaspirated affricates (e. g. aspirated and unaspirated[j]). The articulatory mechanism and the acoustic characteristics of these two types of affricates are not known and may merit investigation.

(2) Voiced-Voiceless. The terms "voiced" and "voiceless" refer to the action of the vocal cords. When a phone is voiced, the vocal cords touch each other and vibrate; when it is voiceless, the vocal cords remain still; they are spread far enough apart that only a slight friction is created by the air rushing through. The following list gives a few examples of pairs of phones which are identical except that one is voiced while the other is voiceless.

Voiced			Voiceless		
Usual Spelling	Phonetic Symbol	Word Containing Phone	Usual Spelling	Phonetic Symbol	Word Containing Phone
th	ð	then	th	þ	thin
v	v	veer	f	f	fear
z	z	use (verb)	s	s	use (noun)
j	j	jeer	ch	c	cheer

The following list shows the contrasts of voiceless-aspirated with voiced unaspirated. The third major column gives examples of voiceless unaspirated stops.

Column 1 Voiced-unaspirated			Column 2 Voiceless-aspirated			Column 3 Voiceless-unaspirated		
Usual Spelling	Phonetic Symbol	Word Containing Phone	Usual Spelling	Phonetic Symbol	Word Containing Phone	Usual Spelling	Phonetic Symbol	Word Containing Phone
g	g	gate	k	k'	Kate	k	k	skate
d	d	dale	t	t'	tale	t	t	stale
b	b	bill	p	p'	pill	p	p	spill

It should be noted that the phones of Column 3 do not contrast (are not used to differentiate words) with the phones of Columns 1 and 2 in the same way that the phones of Column 1 contrast with the phones of Column 2. The phones of Column 3 can occur only after s (at least in English). The phones in Columns 1 and 2 can occur everywhere except after s. The phones of Column 3 have one characteristic of Column 1 and one of Column 2. Phonetically, the p of spill is a compromise between the p of pill and the b of bill.

(3.) The Voiced Nasal Resonances. Voiced nasal resonances occur when nasal passages are open and the vocal cords are vibrating. The nasal cavity sets up resonances which differ from the oral resonances in two major respects. The first formant is weaker for nasal resonance, and an extra formant, frequently referred to as a "nasal formant" is present between the first and second formants. ("First and second formants" is used here to mean the first and second oral formants.)

(4.) Unresolved Classifications. There are some English consonants which have not yet been fitted into the model. These include [h], [r], [y], [w], and vocalic [m], [n], [ŋ], [l]. The factors influencing classification of these problem segments will be discussed in Section 4 of this report, which presents the acoustic data studied on this project. The reader is referred to that section for a resolution of these classifications.

II. VOWELS

This dimension includes all of the phonemically distinct vowels of American English. It also includes all the nasalized vowels and the r-colored vowels; these modified vowels are treated as separate units because their acoustic characteristics are so distinct that it might prove difficult to have the machine recognize them as ordinary vowels.

Nasalized vowels are quite common before nasal consonants in American English. As a matter of fact, it is unusual for an American to say the word can with a vowel which is not nasalized. The difference between a nasalized and purely oral (non-nasalized) vowel is sometimes word-differentiating. In rapid speech, final [t]'s and [n]'s sometimes are omitted, and when this happens the only difference between cat and can (with the meaning "be able") is that cat has a purely oral vowel, while can has a nasalized one.

The r-colored vowels are also quite common in American English. They occur before r. After vowels r is sometimes omitted, and when this happens, the r-coloring may be word-differentiating.

We have two reasons for treating the vowels as a separate dimension.

First, there is a major problem in segmenting a vowel from its adjacent consonant. Second, there is evidence that the articulation (and hence the acoustic characteristics) of a given consonant depend in part on what vowel follows.

The t of tea, for example, is not identical with the t of top. Researchers have found they could not divide a vowel from surrounding consonants without some overlap of the two sounds. For this reason, we must specifically study the combinations of vowels and initial consonants.

A. SEGMENTATION

Until fairly recently, phoneticians did not realize that segmentation was a problem. They worked with very few instruments, and their most important technique was to articulate the sounds they were interested in and observe their own articulatory processes closely. One difficulty with this approach was that when they articulated a sound they were interested in, they sustained the sound far longer than any normal speaker would; hence they gained the impression that the transition from one sound to another is only a tiny fraction as long as its steady-state. When X-ray motion pictures were made, however, everyone who looked at them realized that steady-states were only a small part of the total duration on an utterance.

At the time of the first X-ray movies it was still possible to say that the steady-states, although brief, were the essential part of the speech wave, and the transitions had no function in the perception of speech; later experiments with tape-recorders have shown that this is not the case either. The procedure in these experiments was to erase part of the recording and then note the listener's response to the remainder.

Martin Joos described (Joos 1948 p. 121, 122) an experiment with the syllable tel (from the word hotel). He first cut off the stop-gap, noise burst, and aspiration of t, as well as the first 20 or 30 milliseconds of the voiced portion which Joos considered part of the vowel. When he played this tape to a group of listeners, the largest number said they heard tell, a smaller number said dell, and a few said sell, ell, or hell. Since only a very small number said ell, this means there are some traces of a consonant in the voiced portion. More specifically there are traces of a consonant articulated with the tip of the tongue, since all the consonants the listeners mentioned are tongue-tip consonants except h.

The results of this experiment were quite striking because t is phonetically described as a voiceless aspirated alveolar stop, but the listeners said they heard t in a speech fragment which was voiced, had no stop gap, stop burst, or aspiration, and which had formants like a vowel. This conclusively proved that there are clues about preceding consonants in the actual vowel portions themselves.

The source of these clues can be discovered by studying spectrographic analyses of given syllables within a word. These analyses show that when a vowel is preceded by a consonant, the formants of the vowel (the concentrations of energy at certain frequencies) show a fairly rapid change in frequency at the beginning of the vowel. Such changes are called transitions. When these frequency changes cease, the vowel has reached a steady-state. Research indicates that these transitions vary according to what consonant precedes the vowel. This gives listeners the chance to identify given consonants within actual vowel sounds.

Further research has indicated that the first-formant (lowest frequency) transitions give information about the dimension of resonances useful to our model while the second-formant (second lowest frequency) transitions give information about place of articulation. In all the experiments described below, the experimental material did not contain any consonant clues except transitions and there were no noise burst to indicate release of a stop, yet listeners were able to identify from transitional vowel formants, the presence of a specific consonant.

Among the studies of first-formant transitions of which we are aware are those which have been made at Haskins Laboratories with the Pattern Playback speech synthesizer, which reproduces speech from artificially produced spectrograms. Researchers report that in several experiments in synthesizing speech, when the first formant is kept level, the most "natural" voiced stops (b, d, and g) were produced when the formant was at its lowest frequency. (Delattre, Liberman, and Cooper, 1951). Later experiments (Liberman, Delattre, and Cooper, 1958) showed that when the starting point of the first formant was raised and the start of the formant delayed, listeners reported hearing voiceless stops (p, t, and k). The authors carried out more experiments to separate these two variables and concluded that a rising first formant is a cue to voiced stops, and a time delay in the first formant without rising transition is a cue for voiceless stops. C. G. M. Fant is also studying this aspect of synthetic speech.

Experimenting with second-formant transitions, scientists at Haskins have introduced the concept of the locus, which is the frequency level from which the second-formant transitions of a given consonant are presumed to begin. In synthesizing speech, they report better results if the second-formant transition does not begin at the locus, but simply "points" to it (Delattre, Liberman, and Cooper, 1955). They concluded that the best g is produced with a locus at 3000 cycles, the best d at 18,000 cycles, and the best b at 720 cycles.

Since the locus of a stop is fixed and the frequency of the second formant of a vowel depends on what particular vowel it is, it follows that the transition of a consonant may be rising before some vowels, falling before other vowels and level for one vowel. Thus the transition for [di] is rising, for [dɛ] it is level, for [du] it is falling sharply. For [bi] the transition is sharply rising, and for [bu] it is slightly rising.

All these transitions which give information about consonants occur in what is traditionally considered the vowel portion of the speech stream. Joo's tape-erasing experiment, moreover, showed that both the consonant and the vowel were perceived throughout almost the entire stretch of what is usually considered the vowel portion. As it is virtually impossible to separate a consonant from a following vowel, vowels have been included as a separate dimension in the model to avoid segmenting between them.

B. NON-DISTINCTIVE CONSONANT DIFFERENCE DEPENDING ON FOLLOWING VOWEL.

Very little work has been done on this aspect of consonant vowel combinations, but the available data indicate that frequently a consonant is influenced by the following vowel. Liberman, Delattre, and Cooper report (1952) that the judgements of synthetic stop bursts as p, t, and k depended on the frequency position of the burst in relation to the vowel; this was especially true of p and k. Burst at high frequency were reported as t. Bursts at lower frequencies were reported as k when they were on a level with, or slightly above the second formant of the vowel; otherwise they were reported as p. These data were used by Denes and Fry in the design and construction of their phonetic typewriter.

One aspect of the influence of vowels on preceding consonants which requires further investigation is labialization of a consonant before a rounded vowel such as [u] . In such words as sue and too, many speakers have their lips rounded during most of the consonant articulation; the acoustic characteristics of such labialization are

not known.

III. DURATION

In approaching the problem of the duration of units of speech it is best to emphasize our most basic task; in this study we are trying to break normal speech into patterns that a machine can recognize. A human ear can adjust to variations in regional accents, even to mistakes in pronounciation, without difficulty. A machine must be programmed to isolate each sound from surrounding sounds and to identify it. If two people pronounce a word differently, the human ear makes automatic adjustments; the transcribing machine, however, must be built to compensate for such differences by breaking the word into its component sounds.

To accomplish this difficult task, it is apparent that the classification of each sound within a word must be extremely precise. Past studies of phonetics have tended to concentrate on distinctions which the human ear can identify and which are useful in distinguishing whole words from each other. Often sounds were measured only by ear. These studies are invaluable to machine linguistics because they begin to identify problems which we now must solve. On the other hand, in measuring the duration of sounds of words they seldom needed to break every sound into all its component parts. For our purpose they are not definitive.

It is the problem of dividing each sound into units basic enough to be recognized by a machine that we must now examine in detail. In this problem we face the question of separating each sound in a word from every other sound: when applicable, breaking each individual sound into its beginning (on-glide), middle (steady-state) and conclusion (off-glide); and finally measuring the duration of the sound at each of its stages. (See Figure 8).

These methods of duration measurement discussed in the following sections assume the availability of a reliable formant tracker. At present such a formant-tracker does not exist, but available information requires our assumption that it can be developed. In case it seems extremely difficult or impossible to build a reliable formant-tracker, the following discussion would still be necessary to relate information about how words are pronounced to our common knowledge about how speech is heard. If it is impossible for our machine to identify spoken language by immediate comprehension of formant variations, moreover, it may still be possible to recognize

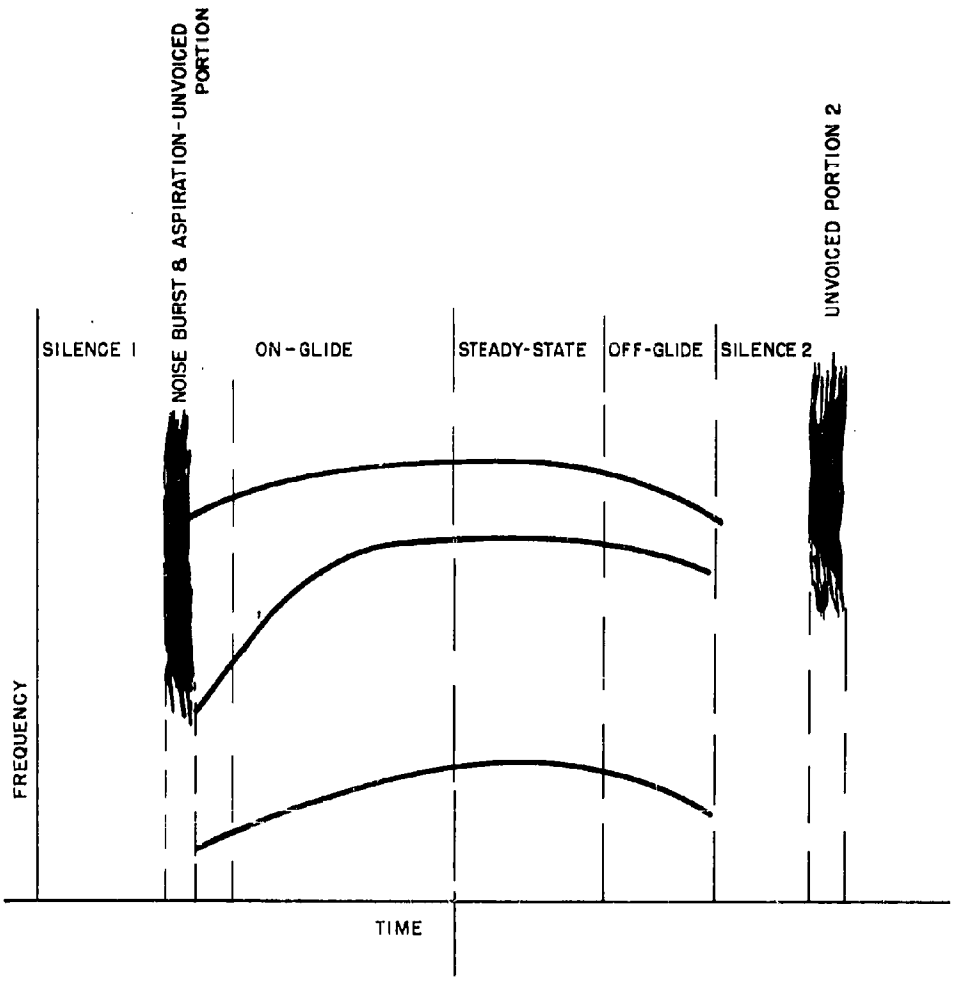
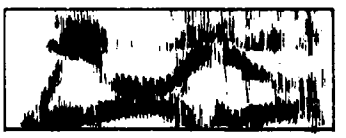


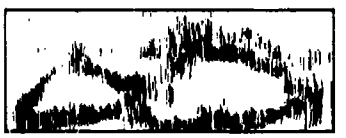
Diagram of onglide, steady-state and offglide portions

Figure 8

General American



English



Southern

Spectrogram of Southern, English, and General American Speech

Figure 9

a word by measuring differences in the duration of its componant sounds by methods that cound be considered later.

Assuming the existence of a formant tracker, this form of measurement would be one way to distinguish between British and Southern accents, for example. The Englishman tends to bite off his words; his on-glide and off-glide are rapid, while his steady-state is comparatively long. The Southerner may take about the same amount of time to make a sound, but he drawls; his on-glide and off-glide are gradual, his steady-state is extremely brief. (See Figure 9) Measuring speech in specified time-units can help to identify these differences and compensate for the quantitative information about formant levels that is expected according to normal standards of the transcribing machine. Phonetic differences in regional speech must be indicated by other criteria, of course.

At times duration is the only method of distinguishing between words, as in "bomb" and "balm". The time durations of on-glides and off-glides are also determining factors in distinguishing semi-vowels from stop consonants.

With the value of duration measurement clearly in mind, we may turn to the problem of utilizing duration in our model. The following subsection will first discuss the work of past transcribing machines and the normalization of duration. Next it will discuss the duration measurements our own model must make. Finally it will summarize the value of phoneticians' past work to our project.

A. Some Considerations on the Normalization of Duration

Any discussion of duration measurement would be pointless without a review of past work which has actually enabled working machines to recognize preselected spoken words. Both automatic digit recognizers and automatic word recognizers are examples of this kind of machine. Although effective within a limited context, such a machine uses methods which are unworkable for a general purpose transcribing machine for reasons described below.

Machines such as those mentioned above can actually recognize given words spoken at different times by different speakers by standardizing or "normalizing" the duration of each of these words. The machine is given one pronounciation (and duration) of the word "nine" for example as its standard for deciding whether a spoken sound is also the word "nine," or whether it is some other word. Every man does not pronounce "nine" with the same time duration. To compensate

for this the machine will proportionately shorten or lengthen the time duration of the sound it hears to conform with its "normalized" pattern of duration for the entire word.

Such standardization of normalization of the duration of words results in proportionate lengthening or shortening of the on-glides, the steady-state, and the off-glides of voiced portions of speech. It has similar effects on the duration of silent portions, of noise bursts, and of periods of aspiration.

The end result of normalization is that the digit recording machine can recognize the standard pattern of voiced parts of a word (parts using the vocal flaps, such as m) as they may be spoken by several speakers from the same locality and with a similar dialect. The duration of unvoiced parts of sound (those not using the vocal flaps, such as "s") presents a different problem, however. The durations of certain unvoiced sounds such as "sh" can vary considerably with each individual speaker and each regional dialect.

Given two conditions a transcribing machine may be able to recognize words by standardizing their duration. The first condition is that voiced sounds be a prominent part of the pronounciation. More important, the machine must have a limited vocabulary, such as ten digits. A third possible condition could be that speakers have the same regional accent.

Today's transcribing machines are able to ignore small differences in pronounciation precisely because they have limited vocabularies. To identify a word, they compare the patterns of pronounciation of whole words rather than of their component sounds. The words in a machine's vocabulary represent extremely diverse wavetones. When only a small number of dissimilar patterns need to be identified, the task is simple. It is made even more easy by the fact that words are spoken separately into digit recording machines rather than slurred together as they would be in normal speech.

The general purpose speech recorder cannot work under these limitations. It must have a large vocabulary. With a large vocabulary it must be able to distinguish between very similar patterns of pronunciation. It will not be transcribing the slow precise tones of telephone operators enunciating a long-distance number; rather it will be recording language spoken at its normal rapid rate. Finally, when such similar words as "three" and "through" are included in the vocabulary, a general transcribing machine cannot be limited by inability to recognize units smaller than simple word-patterns.

A general purpose recognizer must consider the sum of the sounds as they combine to form a whole word, not simply the sound of the whole word. The following section will thus consider the importance of measuring the duration of separate sounds as they naturally occur within a word.

B. The Importance of Duration Measurements to Speech Recognition

In English there are many distinctions between individual sounds that can be identified accurately only be measuring the duration of a given phone. Notable examples are summarized below.

1. Some words are identical except for vowel duration (balm, bomb).
2. In English the phones of a stressed syllable are longer than those of the same syllable when it is not stressed (the by itself vs. the apple).
3. One difference between voiced and voiceless consonants in English is that a vowel preceding a voiced consonant is longer than the same vowel preceding a voiceless consonant (bead, beat)
4. Some pairs of similar consonants have duration differences which serve to identify them. Rapid is distinguished from rabid because the stop-gap of [p] is longer than that of [b]. One difference between English s and z is that s is longer than z .
5. Transitions and semivowels are also differentiated by duration. A semivowel is longer than a transition (cue, coo).
6. There is at least one Southern dialect in which the transitions from phone to phone have a very long duration, and the steady-states are very short. In any dialect where this is the case, it seems probable that the steady-state does not reach the frequency level it might have attained if the transition had been more rapid. It may be necessary to instruct the machine to compensate for this whenever such a brief steady-state occurs.
7. A non-final nasal is much shorter than a final nasal, and if a final nasal has the duration of a non-final nasal, listeners will report hearing a voiceless stop after it.
8. The American English vowels I (pit), ɛ (pet) ə (putt), ʊ (put) differ from i (peat) æ (pat), a (pot) , ɔ (bought), u (boot) in that the former are shorter and they also have a longer offglide relative to the steady state.

We will next consider how traditional phonetics has analyzed the speech wave pattern to solve such problems of identification, how our work differs from most past work, and how our work can utilize the results of other studies.

For traditional phonetics duration has been important primarily when it served to distinguish one word or phrase from another. Such distinctions are phonemic (word-differentiating). In some American dialects, for instance, the words balm and bomb are identical save for the duration of their vowels; because it distinguishes the two words, the difference in duration is phonemic and of greater practical significance to phoneticians.

The problem of identifying sound by ear has tended to confine phonetic studies to problems of phonemic duration; this is unfortunate for our present investigation. Although there is a considerable degree of difference in length between a very long and a very short phone, for example, there is likely to be only a minute difference in duration between two similar phones, as in "buck" and "duck". There are many such small differences in duration, leading gradually from phones of very short duration to those of very long duration. Only when length is phonemic are phones likely to fall easily into "long" and "short" categories. Phoneticians listening to sounds do not identify them in terms of precise time units, moreover. Instead, they identify long and short sounds in terms of how they are heard in their phonetic context. The vowel sounds in biz (as in show-biz) and beat are probably of the same duration, but the i in biz seems to be shorter. For our model most of the finer distinctions are important in analyzing a sound because the model must measure them.

In addition to the distinction between sounds which are phonemically long and those which are phonemically short, linguists have also paid special attention to length variations which serve to characterize word boundaries. The most commonly cited example for this is the contrast between the phrases a nice man and an ice man. According to an analysis which is widely accepted among linguists the phonetic differences between these two phrases is that the n in an ice man is longer and more drawn out. Acoustically, this is not true (see the discussion on nasals) but the whole problem is extremely complex and many linguists continue to use the old description because no clearcut new description has been proposed.

Aside from phonemic distinctions which have just been discussed, the only other functions of duration much discussed by linguists have

been the greater duration of a stressed syllable and the greater duration of a final syllable. In both cases, however, the duration variations are also accompanied by pitch and intensity variations. The problem is thus not purely one of duration, and requires an exhaustive study not limited to the purposes of this particular subsection.

We may now turn to the actual work already accomplished in measuring duration. We have already discussed how to divide sounds so that a machine can measure them. This problem again rises when we try to assess the effects of duration variations, since it is frequently necessary and extremely arduous to decide the specific point at which a particular phone begins or ends.

Many researchers on duration have used criteria for segmentation that are different from those we are considering. Hence much of the work we are now going to discuss in this subsection may have limited application to our model. We will first review those explorers whose work, although not directly useful to our study, lights the way to further research. Next we will consider more recent studies which break duration measurements into the same units our model plans to use.

One of the early and more comprehensive studies of English phone durations was published in 1903 by Ernest A. Meyer. Meyer used a rubber mouthpiece to record the air-pressure variations in his subject's breath while they spoke. He measured the sound durations from these air-pressure records. He also measured the transitions from one sound to another by machanically recording the lip movements of the speaker. Those parts of the speech process which showed rapid lip movement is one direction he called glides.

Meyer's equipment was quite simple, but it is worth noting that several of his conclusions have been supported by more recent research. The experiments of Denes, Lehiste and Peterson, and Sharf described below confirm Meyer's statements that a vowel before a tense (voiceless oral) consonant is shorter than before a lax (voiced oral) consonant. Dene's work also confirms Meyer's observation that a tense consonant is longer than a lax one.

Nevertheless, several facts about Meyer's work limit its value for machine linguistics. Meyer had only two informants; both spoke standard British English. Some of Meyer's results may simply reflect the idiosyncrasies of his informants. Moreover, statements about British English as it was spoken sixty years ago do not necessarily hold true for American English today. Still another drawback is that the material used for this study consisted of one-and-two-syllable words;

this means that it deals primarily with stressed or "accented" syllables. A fourth drawback is that Meyer used the terms "tense" and "lax" without ever clearly defining them. It is possible to discover from the text which phones fall into each category, but the identifying characteristics of the various categories are never described. (Meyer's terminology and conclusions are presented in Appendix D.)

Agreeing with previous observations by Meyer, Lehiste and Peterson (1960) report variations in vowel duration which depend on the following consonant. They add that the relative durations of onglide, steady-state, and offglide remain constant. If this is true of Lehiste and Peterson's data, it is probably also true of Meyer's data. This would seem to be applicable only when measuring accents of people with similar dialects, however. Lehiste and Peterson also agree that a vowel before a voiced consonant (one using the vocal flaps, as in bag) is longer than before a voiceless one (not using vocal flaps, as in pack) and that fricatives (s, sh, z, zh, f, v, th, h) lengthen the preceding vowel. Lehiste and Peterson agree with Meyer that no definite statement can be made about the effect of an initial consonant on the following vowel.

They disagree with Meyer about the effect of nasals on the preceding vowel. Meyer says the vowel is shortened while Lehiste and Peterson say it is lengthened. Whenever there is such disagreement, Lehiste and Peterson's results may be more interesting to us because their informants spoke the dialect we are studying.

Investigation by Donald Sharf (1962) suggests the importance of the relationship between some vowels and the duration of their following consonants. Sharf recorded word pairs such as catty-caddy, tacking-tagging, and napping-nabbing, which were identical except for the voicing or voicelessness of the stop consonant between vowels, or "intervocalic stop." He reported that he measured the relative duration of the vowels before the different stops, but he did not say what criteria he used to make the segmentation between consonant and vowel. Since there is considerable overlap, this is a serious omission.

Sharf arrived at the following results. The proportionate duration of vowels before p, or b is 3:4; the duration before k, g is 4:5; the average duration of a vowel before d is .9cs longer than before t. Since this experiment did not manipulate durations, but only calculated them, we still have no evidence that the length of the preceding vowel affects the perception of a stop as voiced or voiceless. When Sharf's work and that of Denes (considered below under sibilants) are compared, however, this seems possible and worth investigating.

The preceding survey is valuable mainly for the lines of

investigation it suggests, rather than for its direct bearing on our model. The following studies are of primary technical importance in defining the different aspects of duration which our model plans to incorporate.

Researchers have used several different methods to investigate duration. Perhaps the most valuable series of studies for our report is that of Haskins Laboratories. Using Pattern Playback, Haskins has made it possible to produce artificial sounds and alter sounds by varying one particular detail of the speech wave pattern while keeping others constant. By such variations it is possible to identify phonetically significant aspects in the duration of the onglide.

A second method of experimentation is to record sounds on magnetic tape. Significant work in this field is that of Leigh Lisker, who was able to change the sound of words by splicing taped sounds and varying the duration of silence after stop consonants; of P. Denes, who used similar methods with s and z; and of Richard Harrell who played taped sounds backwards to check the duration of nasals. An additional important method of measuring duration is to analyze sound spectrographs and compare them; principal workers in this field are Ilse Lehiste and Eli Fischer-Jorgensen. Specific discussions of experiments that have supported these conclusions can be found in Appendices D and E.

The outline given here merely summarizes various aspects of duration measurements.

1) Duration of Nasals. The relative duration of initial and final nasals is important in distinguishing between such sounds as bum and bump. Assuming both syllables receive the same stress, evidence indicates that final nasals are longer, but there is some dispute about this.

2) Relative Duration of the Onglide, Steady-State and Offglide of Liquids and Semi-Vowels. Experiments with the Pattern Playback indicate that r, y, and w each have onglides, offglides, and steady-states of proportionately equal duration. The sound of l is most easily distinguished by listeners when the first formant transition is very short.

3) Duration of Spirants and Sibilants. Available research by P. Denes and Ernst Meyer suggests that variations in the duration of spirants and sibilants may be more important than voicing in distinguishing between such words as the noun use and the verb use.

4) Duration of Stop Closures. The listener's ability to distinguish between p and b, as in rapid and rabid may depend on the duration of the stop closure between vowels in English according to tape-recorder experiments made by Leigh Lisker.

5) Duration of Stop-Bursts. The duration of noise bursts may be the same for all stops in English, but there is little definite information on the subject to confirm or deny this.

6) Duration Between the Stop Burst and the Beginning of the First Formant. In this category duration measurement is particularly helpful in distinguishing between voiced and voiceless stops at the beginning of a word so that it may be possible, for example for a transcribing machine to tell the difference between bah and pa.

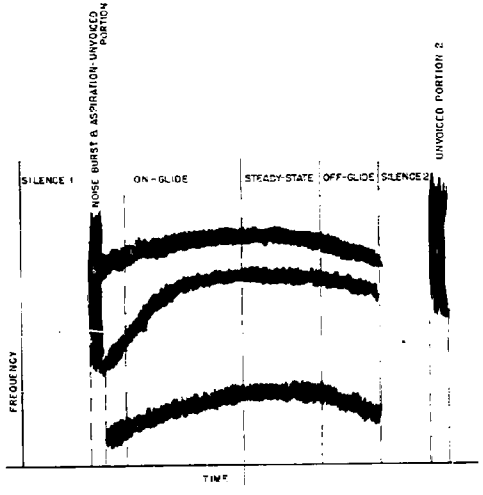
7) Duration of Vowel Onglide, Offglide and Steady-State. These three aspects of duration are closely related to each other, although each will be measured separately in our model. As indicated in the introduction, their relationship to each other may be particularly valuable in helping a transcribing machine to recognize both a Georgia drawl and a Yankee stammer.

Experiments have shown that when the duration of the onglide is disproportionately long, a semi-vowel tends to be heard, after certain consonants. Thus bat becomes bwat. This also explains why Virginians who drawl say gyarden instead of garden. The Southern onglide and offglide are disproportionately long compared to the steady-state according to spectrograms made of Southern speech. There is some indication that a transcribing machine may have to be programmed to compensate for such variations from standard American speech patterns.

C. Methods of Indicating Duration in Our Model

The above data, important in themselves, are also the necessary prelude to the methods of indicating duration we expect at present to include in our model. Having identified the formants of separate sounds of a word according to a set criterion, it is then possible to attack the problems of individual idiosyncrasies and regional accents that still need to be solved.

In Figure 10 we have taken the word pit and divided its sounds according to our proposed method for indicating duration. The upper sketch represents the word as it would appear on a spectrograph. Divisions marked by dotted lines indicate the portions of the word we propose to measure as separate units of duration.



DURATION OF SILENCE 1 — $t_{silence\ 1}$
UNVOICED PORTION 1 — $t_{unvoiced\ 1}$
ONGLIDE — $t_{onglide}$
STEADY STATE — $t_{steady\text{-}state}$
OFFGLIDE — $t_{offglide}$
SILENCE 2 — $t_{silence\ 2}$
UNVOICED PORTION — $t_{unvoiced\ 2}$

pit

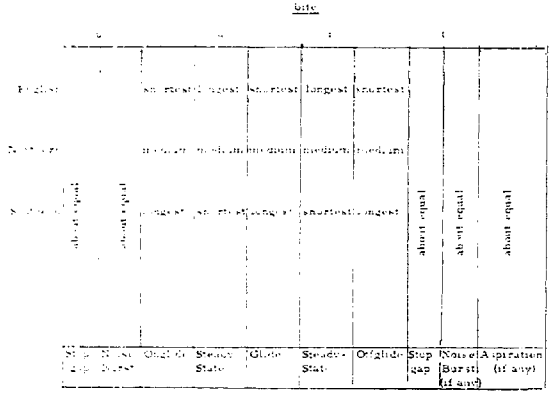


Figure 11

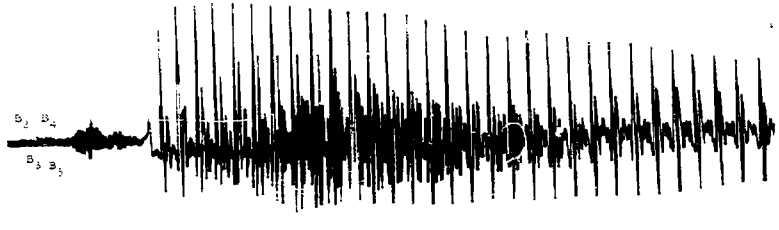


Figure 12

Directly below this diagram we have drawn and labelled the duration units we would use to indicate these separate portions of pit. Most of the divisions will already be familiar from the previous section. Divisions one and six represent the brief silence that precedes the sound of a stop consonant, while divisions two and seven measure the duration of the unvoiced portions. For convenience each division is labeled according to its auditory function, with t representing the amount of time it takes to speak each separate portion of the word. The offglide thus represents the measured time of the offglide. The practical application of this form of measurement is evident when we consider the difference in regional pronunciations of words such as tight, which can sound soft, normal, or clipped depending on whether it is spoken by a Richmond Virginian, mid-Westerner, or announced for the British Broadcasting Corporation. Figure 11 represents the comparative duration variations of these three accents.

In our model, it should be emphasized, such variations will be measured in time units. It should also be noted that some extreme Southern accents make phonetic changes in the vowels of tight which are not indicated in this simple duration measurement.

Normalization and variations in formant levels for pronouncing the same word are additional problems that merit attention. Dividing a sound into its parts, a transcribing machine must still be able to account for variations in the time it takes to hear those parts. Spectrographs of Southern speech also indicate that because of their long glides, speakers do not actually reach the levels of formants normally associated with each identifiable sound. (Figure 9). A transcribing machine must be able to recognize that such variations may still be included in its definition of a sound.

We have discussed how different speakers might pronounce the same word with different durations. Equally important for formant movements and durations are the combination of speech sounds within a word and the order in which they occur. This phase of duration will be considered in our studies of the rules of euphonic combination in Section 3.

IV INTENSITY

Intensity of an acoustic wave is defined as the average rate of flow of energy through a unit area normal to the direction of wave propogation. For wave forms of speech such as illustrated in Figure 12 the intensity in front of the speaker's lips can be defined as:

$$\text{Intensity} = \frac{k}{(t_2-t_1)} \int_{t_1}^{t_2} [a(t)]^2 \quad \text{ergs per second per square centimeter--(2.1)}$$

$$= \frac{10^{-7}k}{(t_2-t_1)} \int_{t_1}^{t_2} [a(t)]^2 \quad \text{watts per square centimeter---------(2.2)}$$

Where:

t_1 = time at the start of the interval of time during which the average acoustic energy in the waveform of speech is to be measured.

t_2 = time at the end of the above mentioned time interval.

$a(t)$ = the amplitude of the sound pressure (or that of volume velocity or that of particle velocity) of the speech wave in front of the speaker's lips as a function of time. $a(t)$ is usually expressed in dynes per square centimeter.

k = constant determined by the physical characteristics of the medium in which the speech waves propogate. For air at a pressure of 781 millimeters of mercury and at 20 degrees c, 1/k-41.1 dynes seconds per square centimeter.

Intensity as defined in equations 2.1 and 2.2 can also be considered as the energy contained in a column p centimeters long that would pass through the unit area in (t_2-t_1) seconds, as:

$$\text{Intensity} = \frac{[\text{Energy in a column p centimeters long}]}{p} \quad c \text{ ergs per second per square centimeter} \quad (2.3)$$

where: $[t_2-t_1] = \frac{p}{c}$ seconds----------(2.4)

and c = velocity of sound in the medium in centimeters per second.

If an acoustic wave were to remain unchanged in its time amplitude characteristics, and if its peak amplitudes were to attain a constant level, over time intervals longer than a second, then its intensity over a one second interval would be the energy in the wave during that second.

For waveforms of normal conversational speech, the uniformity of time amplitude characteristics cannot be retained for periods of time as long as a second; nor can its peak amplitudes attain a constant level during such a long time interval, as illustrated by Figure 12. However, equations(2. 1) and (2. 3) can still be used for identifying intensity over a smaller interval of time than a second.

When such an interval of time becomes extremely small (i.e., in the limiting case when $t_2 - t_1 \longrightarrow 0$) the equations referred to above indicate instantaneous power in the wave. Such power, however, varies very rapidly with time within any period of fundamental vocal cord frequency -- as is obvious in regions A_1, A_2, A_3, A_4 in Figure 12. However, instantaneous power may be a suitable representation of intensity in regions B_1, B_2, B_3, B_4 in Equation 21 wherein the amplitude is relatively constant. In such a case, though, one could also consider measuring energy over the time interval from B_1 to B_5 and obtaining a value for intensity from equation (2. 3).

A similar approach can be considered for the voiced portions of speech, as in region A of Figure 12. For such a measurement of intensity, the period over which energy is to be averaged is the same as that of the fundamental voicing frequency.

For accuracy in such measurement of intensity during the voiced portions of speech, it is essential to measure the fundamental voicing frequency quite accurately. Moreover, the integration of $[a(t)]^2$ would remain unchanged over different voicing periods if the amplitudes of their peaks were constant and if the time amplitude characteristics of the speech wave remained unchanged. Such a situation can arise only when the spectral density of waves is relatively stable, when the voicing pulses are of constant amplitude and when the integration is performed over a complete voicing period.

When the three conditions just mentioned are not met, the value of the integral of $[a(t)]^2$ is known to vary, as readily observable in the fluctuation of the needle on the VU meters used for monitoring speech for recording purposes or for broadcasting purposes. That such a fluctuation of the VU meter indication is due to variation of the loudness of the speech is commonly recognized. The three conditions mentioned above present additional reasons for such fluctuation, as illustrated by considering the calculations of intensity of speech in regions A, B, C, D, in Figure 12.

Moreover, the level of intensity indicated by the VU meter or sound intensity meter may not necessarily reflect the effort required

for generation of a selected portion of the speech wave. For example, a loud enunciation will rarely bring the amplitude levels of unvoiced portions of speech to that of the vowel sound of less loud speech sounds. Such a difference in the level of intensity arises from the mode of articulation and from resonances used.

For reasons mentioned above, it seems advisable to consider alternate methods for indication of intensity. The methods we are considering take notice of the following characteristics of speech waves.

1) Speech waves consist of unvoiced portions of relatively low amplitude and of voiced portions of relatively high amplitude.

2) The amplitudes of unvoiced portions of speech are characterized, in broad terms, by the manner of their articulation, e. g., the waveforms of s have more amplitude than those of f.

3) The amplitude of voiced portions of speech also vary according to articulatory information, e. g. the amplitudes of nasals are usually smaller than those of the vowel sounds.

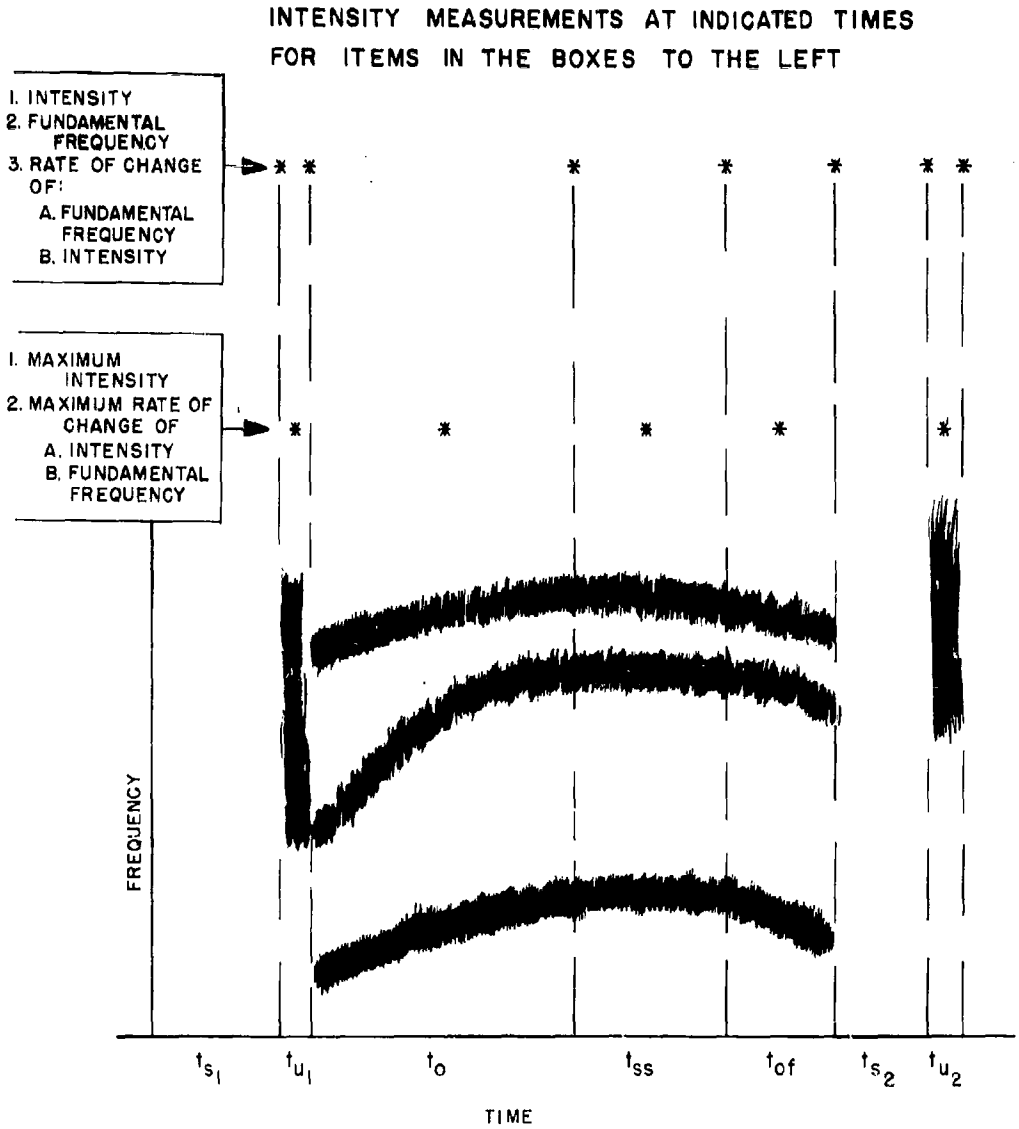
4) Following the articulation of consonants, the amplitude of the succeeding vowels tend to build up to a peak level and then these amplitudes decay before the end of the vowel enunciation.

5) The rate of build up of vowel amplitudes seems to be usually more rapid than their rate of decay.

6) Most unvoiced consonant sounds do not show such time-amplitude characteristics - notable exceptions being the waveforms of ch that show several regions of increased amplitude.

7) The variation of the amplitudes of waveforms of vowel merit investigation for establishing:
   a) their relation to the transients on spectrograms
   b) their acoustic correlates with emphasis or accent or with linguistic stress
   c) their ability to identify stress on the consonants that precede or ones that follow the vowel sounds with these variations

While information regarding the above aspects can be obtained from the "mingograms" of speech, published by G. G. M. Fant; additional stud, of the time-amplitude plots that can be made with instruments having a wider frequency response seem called for.



(FOR INFORMATION ON TIME INTERVALS SEE FIGURE 10)

Representation of Speech
in Phonetic Symbols
Figure 13

For establishing a measure of intensity three methods are under consideration:

1) Study of the spectral density distribution at various time intervals.

2) Measurement of amplitudes of the voicing pulses.

3) Measurement of
   i) radiated voicing pulses

   ii) envelopes of radiated unvoiced portions of speech

Without discussing in detail the relative merits of the above methods, the method (3) seems to be most easy to implement. However, the validity of dat. obtained by any of these measurements, in light of the special characteristics of the speech mentioned before, requires additional investigation.

Assuming that such a measure were developed, and its measurements related to the components of enunciation, as discussed under duration, the . pr sentation of speech in phonetic symbols would be as illustrate Figure 13.

The objective of such a representation are the retention of information about stress and intonation of speech, as it may be important to interpretation of the meaning of the words recognized.

V FUNDAMENTAL FREQUENCY

Fundamental frequency is the number of times the vocal flaps open and close in a second. Since this number varies from time to time, this frequency can be defined as the reciprocal of the time interval between successive voicing pulses. This is included as a separate dimension particularly because it is helpful in recognition of male and female voices, in understanding of differences in formant levels of these voices, in noting differences in pitch (which distinguish questions from statements), and in recording stressed, or "accented," and unstressed syllables.

A. The Correlation Between Fundamental Frequency Levels and Formant Levels

Not all speakers have the same formant levels, and this must be considered in constructing a model that can recognize all varieties of pronunciation. A woman's speech formants, on the average, are about ten percent higher than a man's, primarily because her vocal tract (the area from above the vocal chords to the lips) is shorter, and also because her vocal flaps are shorter and produce higher fundamental frequency.

By identifying a range of high fundamental frequency and correlating it with a high formant level, a speech recognizer could compensate for deviation from standard male formant levels.

B. High Fundamental Frequency and the Accuracy of Formant Measurements

Because a woman's voice has about twice the fundamental frequency ot a man's it has only half as many harmonics of the fundamental voicing frequency within a given formant band. This lack of harmonics makes any measurement of formant frequency levels less precise; our model recognizes this situation.

C. Pitch and Fundamental Frequency

"Pitch" is a term which refers to a certain aspect of sound perception. The exact relationship between pitch and fundamental frequency has not been clearly defined. We do know that pitch is closely related to fundamental frequency, but it is also related to intensity. Recent studies indicate this to be true, even for complex waves, such as the sound waves of speech. Even for pure sinu - soidal tones, fundamental frequency contributes more to pitch perception than intensity does. It follows that if pitch variations are important in language, an automatic speech recognizer should measure fundamental frequency. In the following discussion we will describe two ways in which pitch is important in language.

D. Pitch as Used in Speech

The meaning of a sentence in English often depends on whether it is heard with a rising or falling pitch. To say "He's coming" with a falling pitch is to make a statement; to say it with a rising pitch is to ask a question. Sentences with such rising or falling pitch have corresponding rising and falling fundamental frequencies; together with intensity, fundamental frequency analysis can help a speech transcriber recognize such vital differences.

E. Fundamental Frequency and Accent

There is some evidence that differences in fundamental frequency serve to distinguish between "accented" and "unaccented" syllables. This can help our model distinguish between such words as the noun subject and the verb subject, for instance.

Dennis Fry (1958) and Dwight Bolinger (1958) report that frequency changes within one syllable, while all others are held to a monotone, have the result that the syllable with frequency variation is perceived as stressed. As it is necessary for our model to make the same distinctions that a native speaker would, this means of identification is particularly important.

SECTION 3: SOUND CHANGE AND THE MULTIDIMENSIONAL MODEL

INTRODUCTION

This section deals with the relation of the various modes of physical articulation of speech, discussed in Section 2, to the processes of phonetic change that occur constantly in languages discussed in Section 1. We assume that among the constant phonetic variations that occur within any language there exists a definite order which may be accurately defined by a properly inclusive conceptual scheme. Such a scheme is suggested in part by a careful examination of the phonetic variations within English itself and in part by a broader survey of consistent phonetic changes that have taken place in other languages.

The perils of over-reliance on the evidence of phonetic changes in languages should of course be emphasized; in most cases our sources are limited to written text, and deductive reasoning based on available data must often serve in place of an actual knowledge of how sound change occurred. Nevertheless the identification of such changes that have taken place can serve as a guide to further orderly analysis of sound changes that may occur in English during rapid speech. As a possible means to such orderly analysis we suggest the concept of well-defined planes of articulation that vary according to the physical modes of speech production considered previously.

We will first discuss aspects of the problems which sound change within a given language can raise for our model. We will than evaluate evidence and conceptual approaches helpful in determining sound changes, briefly sketch theoretical causes of sound change, then consider how it may be possible for our model to represent the sound changes or variations that may take place in English during rapid speech. Itemized data are included in appendices, particularly as they relate to observed rules of phonetic change, here tabulated for the first time with the aim of integrating phonetic, phonemic, genetive linguistic and acoustic aspects of speech.

Previously we have considered how it may be possible to identify sounds as they are articulated individually or within relatively simple and isolated word units. As the words or the combinations of words in normal speech become more complex, so also do problems of recognition. Particularly important is the identification of slurs and dropped sounds that occur both within and between individual words.

It is common observation, for example, that such phrases as seemed to are condensed into seemto in normal conversation, similarly the t almost disappears from rents. Such run-on sounds occur as a

continuous wave pattern on a spectrograph, moreover, and the identification of individual word units within this pattern requires that our model be thoroughly familiar with possible phonetic variations that may take place because of rapid speech or individual and regional idosyncrasies.

This requirement suggests three specific needs - an orderly listing of various possible phonetic combinations in which "merging" of sounds takes place, a comprehensive conceptual scheme for defining the boundaries between individual sounds, and logical method for making arbitrary divisions between word units. Procedurally our first problem is to define clearly for ourselves the distinction between phonetic change and phonetic variation; once this has been done, past phonetic changes in Indo-European languages provide the best approach to identifying various phonetic variations or changes that take place today in English.

In the following sections we assume that phonetic variations in modern English actually duplicate phonetic changes that have taken place in other languages. This premise is plausible because all Indo-European languages utilize the same physical modes of production. Our implicit assumption is that phonetic changes are governed by an identifiable set of rules based on physical means of production.

One possible way of deriving a table of phonetic combinations for all languages thus might be to make an intensive study of English; considering the vast scope of such a task, it seems more convenient to apply to English the available evidence on the phonetic changes that have taken place in the past. It may thus be feasible to arrange our model with sections of reference devoted to special phonetic variations and to places where a sound drop-out is likely to occur. Once the various phonetic changes in English have been ordered, moreover it becomes easier to devise some means for arbitrarily separating the acoustically undifferentiated words of normal conversation.

I THE NATURE OF SOUND CHANGES

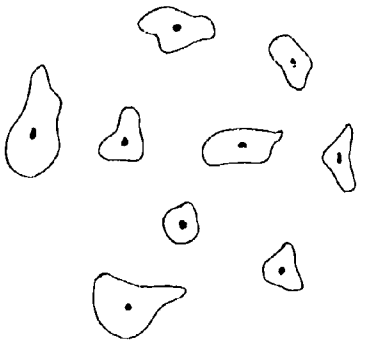
Sound change is a gradual and continual process which takes place in all languages. It is so gradual that the speakers of the language very seldom notice that any change is taking place. Occasionally they notice that the speech of the oldest members of the community differs from that of the children, but they attribute the difference to the effects of aging rather than to changes which have taken place in the language since the oldest inhabitants first learned it. The oldest inhabitants themselves do not realize that their speech has changed since they were

children. They may be conscious of the "modernisms" in the speech of the younger people, but they do not realize that their own speech also contains recently-acquired modernisms.

One reason people fail to notice definite sound changes from one generation to the next is that they cannot distinguish such change from random sound variation. As we have previously emphasized, two different pronunciations of the same linguistic unit are seldom identical. When the word cat is spoken twice, for example, each phonetic unit - the k, the ae, and the t - will probably differ slightly. In analyzing sound we face the conceptual problem of distinguishing such random variations from changes that gradually alter the phonetic structure of a language.

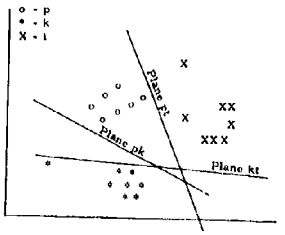
As an aid in solving this problem most phoneticians assume that sounds can be mapped in definite space with defined boundaries for each phone class. Such mapping is primarily a conceptual tool for explaining our tendencies of classifying phones. Traditionally phoneticians have conceived of sounds as units clustered around a specific norm or center of gravity which lies in the center of the area in which varied pronunciations of a given sound are most likely to occur (see Figure 14). Some linguists describe sound change as simply a shift in this center of gravity. While this explanation gives a picture of how sounds shift, it fails to give adequate opportunity for analyzing the physical nature of that shift, nor does it provide the necessary frame of reference within which an orderly model for speech recognition can operate.

As a more preferable way of representing phone classes we are suggesting a different approach - that of sound grouping within moveable planes of articulation (see Figure 15). In our conceptual plan sounds are not grouped by their distance from a center of gravity; instead they are grouped within specific hyperplanes which define their relations to each other. In the two-dimensional illustration each plane is approximately equidistant from the central cluster of sounds. Any sound that occurs within the boundaries of these planes is simply a sound variation. If the direction of these planes shifts, however, then a sound change takes place. In general, these planes may be related to the many physical choices which a speaker makes in shaping his words, and a shift in these planes may be equated with a physical shift in the way a sound is produced. To indicate how this could be done we include a diagram of the tongue's position against the alveolar ridge as it pronounces t in the word tick (see Figure 16). Today the position of the tongue can vary between point A and point B. We will



Clustering of sound units around a center of gravity.

Figure 14



Planes indicating differences in articulation.

Figure 15

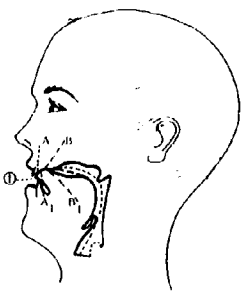


Diagram of Tongue Position for Articulation of t in "tick"

Figure 16

assume, however, that in 1900 the tongue's position could have varied from point A_1 to point B_1, beginning at a point slightly lower than A and never quite reaching as far as point B. In analyzing this change according to our method of representation we would say that one of the planes which defines the sound t has shifted slightly.

It should be emphasized that such a conceptual scheme requires data which can be organized or transformed so that they indicate a definite boundary separating information about one phone class from adjoining classes. Some studies - those of Peterson and Barney's "Control Methods Used in a Study of the Vowels" for example - have reported experiments which show an overlap in the levels of formant one and formant two that are characteristic of vowel sounds. In such cases we must assume either that additional information not investigated in the experiment will make separation possible or that for the vowel sounds no hyperplane can be defined with our present knowledge.

A knowledge of past shifts in planes of articulation is important for automatic speech recognition because they can serve as a basis for predicting which phonetic variations will be favored and which will not. We have many examples of s becoming h, but not the reverse. If a speaker produces a word containing a sound halfway between s and h, we must instruct the machine to look up a word an s in that position, rather than an h.

Most linguists today recognize two types of sound change. These are conditioned and unconditioned. Conditioned sound change differs from unconditioned in that conditioned change takes place under certain circumstances while unconditioned change takes place under all circumstances. During the nineteenth century linguists also believed in sporadic sound change, but this theory was attacked and discarded, because it implied that there is no pattern of sounds in language, and that there is no limit to the number of significantly different sounds which can exist simultaneously in a language. Current linguistic theory is based on the assumption that every language has its own sound pattern and that there is a limited number of phonemes or significantly different sounds.

Of the two types of sound change which are currently recognized, conditioned is the more common because in addition to such cases as "Greek s became h in initial position," this category also includes all cases in which a sound is made more similar to an adjacent one. This special case is called assimilation, and we will discuss it in detail below. Unconditioned change involves such cases as "Proto-Indo-European g became Germanic k." Thus Latin genus and English kin are cognate

(derived from the same Proto-Indo-European word). Information about non-assimilatory sound changes of this type permit us to predict sound-variations which may occur occasionally in any individual's speech. Information about assimilatory changes will permit us to predict the mutations which occur when specific sounds are adjacent to each other.

Assimilatory changes seem to be the result of a strong tendency to simplify the motions of speech articulation. This tendency to simplify one's movements is a powerful cause of sound change. In English the suffix for the past tense was at one time pronounced _d_ in all environments. It is still pronounced that way after an alveolar stop; _tasted_ is an example of this. Elsewhere, the vowel was lost, and if the verb stem ended with a voiceless sound, the _d_ was replaced by _t_. Thus the past tense of _lack_ has a _t_ where the past tense of _lag_ has a _d_. The _d_ of _lacked_ was replaced by _t_ to save the speaker the trouble of changing his vocal flap adjustment during the articulation of the consonant cluster. The process of changing the first of two consonants while leaving the second unchanged is called anticipatory or regressive assimilation. The process of changing the second consonant and not the first is called progressive assimilation. Most phoneticians agree that anticipatory assimilation is the most common type in English. When one word ends with a voiced consonant and the next begins with a voiceless one, the final consonant may become voiceless. Conversely if the final consonant is voiceless and the adjacent initial consonant is voiced, the final consonant may become voiced. Thus the phrase _big pit_ may be pronounced with a _k_ at the end of _big_, and the phrase _thick bit_ may be pronounced with a _g_ at the end of _thick_. The reason for sound changes of this type are quite clear, and we could predict many of them even if we did not have examples from other languages.

In addition to sound changes, there is a special class of linguists changes which are neither gradual nor regular. Some linguists call them sound changes also, but we prefer to reserve this term for the regular gradual process described above. The special category which we are now discussing includes dissimilation, distant assimilation metathesis, and haplology. Dissimilation is the replacement of one sound by another when the original sound occurs twice within a word. The Latin word _peregrinus_ (pilgrims) became _pelegrinus_ when the first _r_ was replaced by _l_. Distant assimilation is essentially the reverse of this process. If two sounds in a word are not similar, one sound will sometimes be replaced by another which is more similar to the remaining sound. The Proto-Indo-European word for _five_ was probably *_penkwe_. (The asterisk indicates that we have not written

records which show this word.) In Pre-Germanic this became *pempe, which in turn became *fimfe. Metathesis is an exchange of position by two sounds within a word or phrase. The Old English word for 'wasp' was waeps. When we compare this with the modern word, we see that the s and the p have changed places. Haplology is the dropping of sound or group of sounds which occur twice within a word. The Latin word nutrix 'feeder, nourisher' comes from an earlier *nutritrix.

All of these linguistic changes can be seen in the slips of the tongue of contemporary speech, but at present we do not plan to include them in our model because they are too unpredictable and because when a speaker makes a slip of this type, he is quite likely to notice it and correct it himself.

There are five different methods for discovering what sound changes have taken place in a given language. They are (1) analysis of the regular phonetic alternations of the language; (2) comparison of the descriptions by different phoneticians, each describing the speech of his own day; (3) examination of written records and poetry; (4) comparison of different modern dialects of the same language; (5) comparison of the written records of ancient languages in order to reconstruct their parent language. Specific discussions of each of the types of reconstruction may be found in Appendix F.

Our criteria for deciding the accuracy of a phonetic reconstruction include the number of different reconstruction techniques which yield this result, the phonetic probability of a given change having taken place, and the number of times this change has been reconstructed for other languages. If several different reconstruction techniques all indicate that a certain change took place, this is very strong evidence that it really did happen that way. For example, the sound-change of the English past-tense suffix is attested by three sources - phonetic descriptions from the eighteenth century, the evidence of spelling, and the morphophonemic alternation of modern English. We have speech-manuals from the eighteenth century manuscripts which spell this suffix as ed, 't, and 'd. Finally, we have the modern morphophonemic alternation, which is most easily explained by the assumption that the suffix was originally ed. Taken together, this evidence leaves no room for doubt.

Phonetic probability is the second criterion for weighing the accuracy of a reconstruction. This criterion is frequently applied while the linguist is working on the reconstruction rather than afterwards. In our analysis of the past-tense suffix, we rejected the assumption that d was the original suffix because it involved the

assumption that *tastd was once the normal form, and this is phonetically improbable. Some linguists have objected to the traditional reconstruction of Proto-Indo-European because it contains voiced aspirated stops b', d', g' but no voiceless ones p', t', k'. The articulation of voiced aspirated stops involves more glottal adjustments than does the articulation of voiceless aspirated stops, and since there was supposedly only one set of aspirated stops in the language, there was no need to go to the extra trouble of making them voiced.

The third criterion for reconstructed sound change is whether the sound change has been independently reconstructed for other languages. The change of d to l has been reconstructed for Latin and Sanskrit; there is also alternation in the Greek dialects between the names Olysseus and Odysseus. If this change should be reconstructed for another language, we would not question it even though at present we do not understand how this change takes place.

From our present evidence we may make two assumptions. The first is that sound change is a gradual and regular process with orderly characteristics of transition. A corollary of this assumption is our belief that an orderly system to deform this change is both possible and necessary. One further aid to the development of such a system would be a workable theory of sound change; previous attempts to develop such a theory are discussed below.

II THE CAUSES OF SOUND CHANGE

Many theories have been proposed as to the causes of sound change, but most of them have been thoroughly discredited. For the purposes of an automatic speech recognizer, however, a valid theory of the causes of sound change would be a valuable aid in predicting sound variations, because it suggests both the probable direction of sound variations as they occur in normal speech and the particular sound variations likely to take place when two given sounds come together, as previously indicated. To be of use in this study such a theory must meet at least three qualifications - it must use physical means of articulation as one of its major criteria for change; it must be sufficiently comprehensive to allow an interplay between the various physical modes of production already discussed; and it must be able to be stated in units comprehensible to our model.

The most commonly proposed theory, is that sound change is a simplification of the articulatory process. This is obviously true of

some cases, but obviously not true of others. The change of [Ink∧m] to [Ink∧m] is a simplification, since it reduces the number of necessary articulatory movements, but the change of Proto-Indo-European t to Germanic p does not seem to be a simplification. Moreover if the change of t to p were a simplification, the change of p back to t in the Scandinavian languages would be the opposite. The theory that all sound change is simplification does not fit the facts.

Other attempts have been made to explain sound change as the result of a change in environment or way of life, but it has always been possible to cite groups of people whose languages did not undergo similar changes though they lived in similar environments with a similar way of life.

Inherent in all our work to date are the assumptions that there is a related order in all language based on physical modes of production, and that such an order may be graphically represented by examining the interaction of these physical modes. A further assumption is that sound change is not random but proceeds along a specific pattern according to cause inherent in the structure of the language itself. Postulating the existence of both an orderly change in speech and an inherent order governed by physical means of production, it appears worthwhile to review data concerning the possible existence of an orderly series of rules for anticipating sound change in European languages, particularly as it relates to these physical means of production.

Historically researchers have accomplished comparatively little definitive work in problems of predicting sound change. Existing theories which assume there is a single cause for change have generally been disproven, when further research disclosed a situation in which the special cause was present, but the expected change did not occur. Andre Martinet, however, assumes that several factors influence sound change: factors inherent in the physical production of language. For this reason, and also for the purpose of obtaining a modern Western linguistic view of sound change, we shall briefly review the work of Martinet. And finally, although it seems to present in an orderly fashion many postulates similar to those on which our own model is based; at the same time it reveals many of the limitations of current linguistic theory when applied to sound change.

Martinet's theory is based on the phonemic theories which have been developed by many different linguists over the past forty years. His unique contribution is to combine these concepts into a theory of sound change. The theory states that many of the causes of the sound changes which take place in any particular language are inherent in the

phonemic pattern of that language, and in the distinctive features-each of which corresponds to one or more of the physical means of production. Thus by carefully examining the pattern, we can suggest which changes are likely and which are not. (A phoneme, as previously discussed, is a class of sounds which do not contrast with each other but which contrast with members of other phonemes. A distinctive feature is a sound quality which, alone or in combination with other distinctive features serve to characterize a phoneme. The differences between Martinet's terminology and that of Jakobson are more fully considered below).

One limitation in applying Martinet's theory to our project is that distinctive features vary from language to language; thus each language analyzed in Martinet's terms must receive special attention to determine precisely what its distinctive features may be. Such a theory may thus be helpful in developing a set of postulates that govern possible sound shifts within a single language.

Figure 17 represents the distinctive features of English considered in terms of Martinet's work according to the units of our model; there are four places of articulation that can serve to distinguish phonemes - labial, dental-alveolar, palatal, and gutteral. The initial consonants of pin, tin , shin , and kin show these different subdivisions of place of articulation which forms one of the three main axis in our model that serve to define how sounds are produced. It is also possible to graph additional distinctive features under resonances and possibly under manner of articulation. In English these contrasts in sound serve to convey different meanings; thus we say that the features are distinctive. On the other hand, the word tin would be recognizable whether the initial consonant were articulated against the teeth, the alveolar ridge, or the palate. In English, therefore, the two positions of articulation for the front part of the tongue are not by themselves distinctive features. In other languages, however, the number of such features may be greater or less; Indian languages, for example, treat the dental and alveolar t's as separate phonemes.

In assuming that the distinctive features in each language can modify sound shifts, Martinet relies on the hypothesis that sound change is likely to occur in those cases when a language already uses all the physical means of articulation necessary to produce a particular sound but lacks the sound itself.

Further assuming that all speech is based on a tension between the need for exact meaning and the desire to minimize exertion in physical articulation, Martinet suggests that such change is more likely to take place within the range of a distinctive group than across a boundary

	Bilabial-Labiodental	Dental-Alveolar	Palatal	Guttural
Stops				
Voiceless				
Unaspirated				
Aspirated	p	t		k
Voiced				
Unaspirated	b	d		g
Aspirated				
Spirants				
Voiceless				
Unaspirated	f	þ		
Aspirated				
Voiced				
Unaspirated	v	ð		
Aspirated				
Nasals				
Voiceless				
Unaspirated				
Aspirated				
Voiced				
Unaspirated	m	n		ŋ
Aspirated				
Sibilants				
Voiceless				
Unaspirated		s	ʃ	
Aspirated				
Voiced				
Unaspirated		z	ʒ	
Aspirated				
Affricates				
Voiceless				
Unaspirated			č	
Aspirated				
Voiced				
Unaspirated			ǰ	
Aspirated				
Laterals				
Voiceless				
Unaspirated				
Aspirated				
Voiced				
Unaspirated		l		
Aspirated				

Figure 17 This chart represents the normal or most common articulation of each American English consonant phoneme.

between distinctive features, since a shift in sound from one distinctive group to another cound make homonyms out of two distinct words. Thus, in the chart of Figure 17 a sound shift might occur between an alveolar t and a palatal t, but it is much more likely to take place between an alveolar t and a dental t, which would share the same distinctive features.

While Martinet's theories may be relevant in suggesting potential sound shifts, there is some question whether they are comprehensive enough to include many of the aspects of speech production necessary to the development of a multi-purpose recognizer. Certainly the extended scope of our model precludes complete assumption of his theories as a basis for organization of sound changes relevant to a general purpose recognizer.

The problems of vowel coloration and observed crossing of the boundaries of distinctive features, in addition to the necessity of re-defining distinctive features for each language also suggest the need for a more general analysis of predictable sound change than presently exists. Such analysis might subsume Martinet's theories as additional data in instructing a general transcribing machine what sound shifts are more likely to occur. While no such analysis for modern English exists in terms which can be used by the multi-dimensional model which we have developed, several factors argue that it may be created. The first is our assumption of an inherent order in all speech directly related to physical means of production. The second is the dominant theory of modern linguistics that sound change is not random; and the third is the tools of genetive phonetic, phonemic, linguistic, and acoustic analysis of sound.

Experiments by Fry and Denes, moreover, indicate that additional investigation is required to correlate the work of phonemic and acoustic analysis with the objectives of our model. Sound experiments by these researchers revealed that their machine differentiation between the k of cook and the t of tick was over 90%, but differentiation between the k of kick and the t of took was less than 75%. This would seem to indicate that k and t can be distinguished by distinctive features, but their acoustic features may not retain such characteristics at all times.

The task of ordering such acoustic data in a comprehensive theory of sound change seems feasible in terms of our model particularly because we must of necessity account for all the physical modes of production which are assumed to provide the basis for directed sound shifts. In terms of our dimensions, for example, it is definitely indicated that the acoustic characteristics of k in kick differ from those of k in cook because the place of articulation is invariably influenced by succeeding vowel sounds. This reemphasizes the im-

portance of consonant-vowel combinations on an automatic speech recognizer.

In the following discussion, we first consider sound change in terms of our multidimensional mode. We then attempt an orderly presentation of certain predictable rules of sound change or euphonic combination. Some of these rules we derived from our studies of historic sound change in Indo-European, Germanic, Old Icelandic, and Celtic languages. Others were suggested by the sandhi rules of Sanskrit, which lend themselves quite readily to a systematic analysis of preferential sound shifts.

Sandhi rules are of a special value because they comprise an integrated chart of euphonic combination developed for a language that in some cases represents phonetic sounds quite precisely. For example, although English makes no distinction between a dental and an alveolar t, Sanskrit has phonetic symbols for both these sounds and regular rules for the euphonic combination of each symbol with other sounds. Other sounds not distinguished in English but specially represented in Sanskrit include visarga vowels and aspirated consonants. Since the Sanskrit alphabet represents an orderly grouping of possible phonetic sounds, and these sounds in turn are based on the physical means of articulation common to all men, it would appear helpful to apply the rules of sandhi to our own development of an orderly method of speech analysis.

III SOUND CHANGES CONSIDERED IN TERMS OF THE DIMENSIONS OF OUR MODEL

The purpose of this section is to examine how sounds may change in contact with other sounds in rapid connected speech; a model for speech transcription must be able to relate these changes to the "purer" patterns of careful speech by recourse to the various subcategories in adjoining columns, rows, or Reimann leaves.

The reader will note that the four main problems to be solved are emphasis, regional dialects, slurring of syllables, and definition of word boundaries. All are particularly important because there is generally no identifiable acoustic break between words in rapid speech. Without means of distinguishing between the sounds of separate words, however, the construction of a general purpose transcriber becomes almost impossible. The following discussion provides for the first time an orderly approach to the solution of this question.

In the following subsections we describe some sound changes which have taken place in the past. This is not a complete list; the

compilation of a complete list would require years. A recent book (Language and History in Early Britain, by Kenneth Jackson) devotes three hundred pages to a concise description of sound changes of the Celtic languages alone. This is simply a sample of the total number of sound changes which have been described by linguists.

A. CHANGES IN MANNER OF ARTICULATION

Changes in manner of articulation involve shifts from one Reimann leaf to another in our charts. This type of change is quite common. It occurs both as an assimilatory change, as when Latin pf became ff, and as a non-assimilatory change, as when Proto-Indo-European p became Germanic f in almost all environments. In modern English a cluster of alveolar stop and [y] frequently becomes an affricate. Thus did you becomes diju and at you becomes æʧu (This involves a change in place of articulation as well as a change in manner). Most changes in manner of articulation which have been described by American phoneticians involve clusters with [y]. We do not know whether this is, in fact, the most common change in manner of articulation or whether it is simply the most conspicuous. Rules for this type of change are included as Appendix H.II.

B. CHANGES IN PLACE OF ARTICULATION

Changes in place of articulation involve shifts from one row to another in our charts. Most changes in place of articulation are assimilatory; they occur only when two consonants are adjacent, but a few, such as the change of Old English final m to n, are non-assimilatory. One assimilatory change is the change of [s] to [ʃ] when it is followed by [ʃ]. This commonly occurs in the word horseshoe, which is usually pronounced [hɔrʃʃu] or [hɔrʃu]. In the Appendix H.I we include a list of changes in place of articulation which have taken place in the past, together with examples of English words and phrases containing the same sound combinations.

C. RESONANCES

Changes in resonances involve shifts from one column to another in our charts. Changes in resonance represent the most common assimilatory sound change. Consonants with different places and manners of articulation occur next to each other in the words of many languages, but it is unusual to have voiced and voiceless consonants adjacent to each other. Almost all languages will permit sound combinations at word boundaries which they will not permit within a word, but the fact that a certain type of sound-combination is uncommon in the words of any language seems to indicate that the same combination may be frequently

modified at word-boundaries also.

There are conditions other than assimilation which cause changes in resonance. In standard German, no word may end with a voiced stop, spirant, or sibilant. The phone becomes voiceless in that position. Thus bunt 'bright' and Bund 'group' are homonyms.

Some changes in resonance are unconditioned; that is, they take place under any circumstances. Thus Proto-Indo-European d became Germanic t. English two and Latin duo are cognate (go back to the same Proto-Indo-European word).

In Appendix H. III we list some of the changes in resonances which have taken place in various languages.

D. HOW SOUNDS DROP OUT

There are two different types of sound drop-out. One is the loss of a sound from a certain position and the other is the loss of a certain sound from any position. The first type includes the dropping of at least one consonant from a consonant cluster and the dropping of certain sounds in final position. The second type includes such cases as the loss of Proto-Indo-European p in almost all positions in the Celtic languages.

The problem of how to treat these drop-outs is a complex one. At present we are attempting to limit the number of positions in the charts from which a sound that is not part of a cluster can drop completely. Thus, at present we assume that [p] does not drop out directly but becomes [h], which becomes [ø], which drops out. In Appendix H. IV we give some examples of sound drop-outs.

E. DURATION

In English, differences in consonant durations are not phonemic (word-differentiating) and the speakers of the language have some freedom to vary these durations. The most common variation is to shorten a long consonant (or a double consonant; we use these two terms as synonyms). Thus red dress may be pronounced with a long or a short [d]. Long consonants which are the result of assimilation or the dropping of an intervening consonant are also subject to shortening: outdoors can be spoken without the [t] and with a long or a short [d]; lasts can be spoken without the [t] and with a long or a short [s].

Conversely, some short consonants may be lengthened under certain circumstances. If a person is counting slowly and rhythmically, he may pronounce eighteen with a long [t] because the preceding words fifteen, sixteen, and seventeen all have consonant clusters in the middle, and if the speaker says eighteen with a short [t], he will break the rhythm.

The problem of length variations due to rhythm is a complex one. Andre Classe has advanced the hypothesis that if there are several strongly accented syllables in an utterance, the speaker tries to vary his tempo so that the time interval from one accented syllable to the next is a constant. Classe calls this equality of time intervals isochronism (Classe, 1939). When the number of intervening syllables is very uneven, the speaker tries to achieve isochronism but does not succeed. If this hypothesis is correct, the duration measurements of all unaccented syllables depend on the positions of the accented syllable. This is a subject which requires extended research.

F. INTENSITY

Intensity of different portions of continuous speech is perceived to be different, except in monotone enunciations. This is a natural phonomenon of control that a speaker exercises to keep his speech from becoming boring.

Such variation in the intensity depends on several factors, some of which are:

1) Emphasis or deemphasis of an utterance for
   a) modification of its meaning
   b) drawing attention to a specific part of it

2) Relative combination of speech sounds that necessitate the emphasis or deemphasis because of natural limitations on production of speech.

The variations of the first type are often controlled by the grammar and syntax of a language, and they do not merit consideration in this part of the study of controlled variations caused by the combination of speech sounds. The variations of the second type merit discussion; but neither is orderly definition of these available, nor can information about these be separated from that for the first type of variations.

An additional difficulty in this field is the need for an acceptable definition of intensity and the interdependence of intensity variation and of variation in fundamental frequency. An orderly study of the effects of intensity would require extensive additional study. In Section 4 we present examples of acoustic data which show the necessity of intensity measurements.

G. FUNDAMENTAL FREQUENCY

Fundamental frequency, like intensity, varies from one portion

of speech to another for reasons that are similar to those mentioned in the previous subsection.

Unlike the measurement of intensity, however, one can find a generally acceptable definition of fundamental frequency that can be used for these studies. However, the information on this subject needs to be further evaluated.

H. GLOTTAL ADJUSTMENTS

The two glottal positions which are commonly used in speech are the positions for voicing and for voicelessness. Besides these, however, there are other glottal positions which are sometimes used and which result in a different spectral pattern. Two of these adjustments are visarga and laryngealization.

During a visarga vowel the vocal flaps vibrate, but they do not touch each other as they do for normal voicing. The resulting vowel has a somewhat breathy quality. In Appendix I we include a detailed description of the visarga vowels.

One theory of the production of a laryngealized sound is that the vocal flaps take on an hourglass shape, and both the front and back halves vibrate while the middle and the ends are relatively still. The resultant sound has a slightly grating quality. Laryngealzation is often used in American speech as a substitute for a drop in fundamental frequency at the end of an utterance.

IV. SOUND CHANGES INVOLVING "PROBLEM" PHONES

A. CHANGES INVOLVING PHONES WITH A SINGLE PLACE OF ARTICULATION

The phones which commonly occur in English but which present certain problems include [y] (you), [r] (right), [h] (how), and the glottal stop [ʔ] (the pause between vowels of uh-oh!). The glottal stop also commonly replaces t in certain words and phrases, such as what was; for this reason it must be included in the model.

There are several separate reasons why these phones are problematic. Y is a semi-vowel corresponding to [i] (the vowel of eat) in manner of articulation save that y functions as a consonant. Although we have classed y with the consonants, its similarity to a vowel creates unresolved problems in describing manner of articulation. In the case of [r] it is possible that the place of articulation of this usually retroflex consonant can vary in American English. The sound [h] is traditionally described as a voiceless vowel, since they receive a separate dimension; it seems necessary to devote a separate set of Reimann leaves to [h]

plus vowel combinations. The problem with the glottal stop resluts from its place of articulation; our model provides for no place of articulation further back than the guttural, but the glottal stop is articulated further back.

In the Appendix H.IV to this section we include a list of rules affecting these phones. The rules for [y], [h], and [r] are likely to be much more complete than the rules for [ʔ], both because the foreign languages from which we have derived our rules have [v], [r], and [h], but no glottal stops; and because the occurances of the glottal stop requires further investigation into the phonetics of American English. Although we doubt that the glottal stop replaces an initial t, for example, this has not been proven. There is also the additional possibility that the glottal stop may replace final t before gutteral (that girl) as well as before a bilabial. In certain New York accents, moreover, the glottal stop also replaces t before l as in little.

B. CHANGES INVOLVING PHONES WITH TWO PLACES OF ARTICULATION

There are two commonly-occuring consonants in American English which have two simultaneous places of articulation. These are [w] as in wit, and [k^w] as in quit. Both of these consonants involve simultaneous gutteral and labial articulation. This combination represents the most common type of simultaneously articulated (co-articulated) consonant, known as labiovelors. Since labiovelars have two different places of articulation, they can also have two manners of articulation. For [k^w] the lips are rounded and open, while the tongue makes a complete closure at the back of the mouth. For [w], there is constriction both at the lips and at the back of the mouth, but there is no complete closure. The rules in Appendix H.V deal with sound changes that have affected labiovelars in other languages.

V. THE RELEVANCE OF SANDHI RULES OF SANSKRIT TO OUR MODEL

In developing a general purpose transcriber the rules of sandhi are particularly valuable because they seem to be an indication of how sounds are produced under conditions of normal continuous speech. A partial reason for this is that Sanskrit grammarians who formulated sandhi rules wished to produce a set of precepts to describe a language as it was currently being spoken. In describing this language, moreover, the grammarians wished to produce phonetic clarity as well as grammatic precision. As a result sandhi rules give particular attention to problems which English grammar (as opposed to the science of phonetics) tends to ignore. Thus, sandhi includes not only rules of euphonic and grammatic

combination within words, but also the phonetic combinations likely to occur when two words come together; such combinations, moreover, are usually expressed in the Sanskrit spelling as well as the rules of grammar.

To the problem of changes caused by coalescence of sounds between words, English phonetics has given relatively little attention, but the importance of being able to identify such changes with our model is apparent. Speakers tend to pronounce their sentences in rhythmic phonetic phrases whose boundaries need not coincide with those of the words involved, a situation that provides the basis for familiar jokes about children who return from church singing songs they learned orally about the three kings of "ory and Tar" (three kings of Orient are. . . .). Such verbal configurations, as previously indicated, are the result of the natural transfer of a sound from the end of one word to the start of a following word whenever the conditions of physical articulation make this easier for the speaker than pronouncing the two words distinctly and the transfer may not interfere with clarity of meaning.

According to our present evidence English phonetics does not seem to analyze fully the problem of transcribing such transfers in an orderly fashion. Sandhi rules, however, distinctly recognize the possible shift caused by the coalescence of a final sound with an initial one, and the sound change so produced is often formally defined. Recognition of such combinations between words is an important aspect of a general purpose transcriber, since the acoustic characteristics of a particular letter may vary considerably depending on preceding and following sounds. r and l, for instance, are extreme examples of this. The following discussion considers particularly those aspects of sandhi that are relevant in expanding our comprehension of sound changes likely to take place between the sounds of separate words, while it provides a rationale for further rules of euphonic combination presented in Appendix II. V. In using Sanskrit rules to help us predict English sound changes, however, we must consider the nature of each rule before deciding how to utilize it.

There are two types of sandhi rules, and while both are helpful in our model, their applicability is different. The first type of rule is the result of conditions which prevail in all human language, and therefore this type is directly applicable. A typical rule of this kind is that when a voiceless consonant and a voiced consonant come together, one is changed so that either both are voiced or both are voiceless. This rule stems from the fact that human beings find it easier to pronounce two voiced or two voiceless consonants together than to change the vocal flap adjustment in the middle of a consonant cluster.

The second type of sandhi rule is the product of historic survival or analogic change. Historic survival is the retention of a sound in a particular phonetic environment after it has been changed or dropped everywhere else. Analogy is the extension of a rule from one group of forms (in this case sounds) to a group of similar forms. There is a sandhi rule to the effect that before certain voiceless stops n is changed to anusvara (nasalized vowel) and a sibilant is inserted between the anusvara and the following stop. This rule represents both an historic survival and an analogic change. At an earlier stage of the language, there was a very large group of words ending with ns and a smaller group ending with n. As the result of a sound change the s was lost from the ns words everywhere except before certain stop consonants. This meant that in most environments the ns words were identical with the n words, but in certain cases the ns words had an s, while the n words did not. Gradually people forgot the difference, and since the ns group was larger, the rule which applied to that group was extended by analogy to the n group.

We would not expect this rule to apply to English directly because English has not undergone the sound change necessary for the historic survival. It does suggest, however, that it is worthwhile to see whether the same processes might have acted on some other English sounds, and we do have one example of that: the so-called "intrusive r" of New England speech.

In some New England dialects final r was lost after vowels everywhere except when the next word started with a vowel. The result was that the pronunciation of the word deer was different in the phrases deer walks and deer is. The pronunciation of deer before a consonant was identical with the pronunciation of the last part of idea in all phonetic environments. Gradually the rule about r spread from deer is to idear is.

A careful examination of all historic and analogic sandhi rules together with a study of the history of English sounds should probably serve to point out more such rules which apply to English.

In Appendix H, V we have separated the rules which we believe to be historic and analogic from those which we believe to be phonetic. The other rules of sound change in Appendix H are all phonetic.

The first particular contribution sandhi may make for our model is to furnish an orderly guide for the coalescence of the final consonant of one word with an initial vowel of the following word. In Sanskrit the words aham ādityāir are written according to syllables, as follows:

a ha mā di tyāir. The final consonant of aham goes with the initial vowel of ādityāir. The combining of a final consonant with an initial vowel is quite common in English speech. Most people distinguish an aim from a name only when they are being extra careful. This sound change must be recognized in our model, since we treat consonant-vowel combinations as a unit.

A second potential contribution of sandhi rules is to delineate possible conflicts and sound changes that can occur in the articulation of one or more consonants which are part of a consonant cluster. The phrase sit down is likely to become sidown in normal speech, as has already been pointed out. On the other hand, the same drop-out of a voiceless stop next to a voiced stop with similar place of articulation is not likely to occur in the phrase look good, because a drop-out of the k might destroy clarity of articulation and consequently clarity of meaning.

With further research into the effects of combining different physical articulations, we may find it possible to utilize sandhi rules in setting up a relatively complete set of instructions to inform our model exactly when and how changes will take place.

A further advantage in applying sandhi rules to our model is that they point out the existence in English of certain sounds which had hitherto been considered peculiar to Sanskrit. One of these sounds is the visarga, which is a vowel-like sound similar to h. In Sanskrit visarga occurs under certain circumstances as a substitute for final s or r. In English we have observed it in New England speech at the end of the words car, law, and December (see P. Denes, Computer Processing of Acoustic and Linguistic Information in Automatic Speech Recognition, Contract No. AF 61(514)-1176, March, 1962, University College London, England, Fig. 9(a), page 27), and before the final k of park. If the Sanskrit rule applies to English, it covers only car and December, since the r of park is not final and law has no r. It is possible that in English visarga can occur with any final vowel. (In the New England dialects car ends with a vowel, since there is no final r.) This rule would cover car, December, and law, but not park. It may be that visarga occurs in park only when the complete closure of the stop is held so long that the vowel is like a final vowel.

As outlined above and more specifically tabulated in Appendix H. V, sandhi rules present a highly useful guide in the development of an organized set of data on which to base instructions about sound change for our model. With an orderly representation of phonemes and an orderly representation of the rules of sound change, it can become possible to present rules for the speech of selected languages organized

so that the rules may be suitable for mathematical treatment or for computer analysis.

VI. REPRESENTATION OF RULES FOR EUPHONIC COMBINATION

In past reports we have indicated that in the course of normal speech many sound changes occur. In verbal form we have presented long lists of such changes or possible changes. If, however, such data is to be available to a machine, a mathematical (i. e., symbolic) representation is necessary. The rules listed in Appendix H. VI present such a digest of the data already presented.

The form these rules take is that of "ordered," speech-environmental," rules which may be considered from a "phonomenological" point of view. These terms require explanation. Ordering of rules means that some rules have priority over others. If, for instance, there were a rule s becomes z ordered before a rule that st becomes s, the second rule would never take effect in any of those situations where the first rule applied.

Speech-environmental rules are based on the idea that the changes which occur in any particular sound (phone) are determined wholly by the nature of the few surrounding phones plus the special characteristics of the speaker. Such a view is supported by our contention that the changes which occur during speech are determined by the mechanical nature of the speech-producing machanism. The characteristics of the speaker determine where he takes care in enunciation, and so in particular take account of those phonemic distinctions which are made in his language.

Phenomenological rules are those which describe an outcome for the physical articulation of speech. They are to be contrasted with probabilistic and random rules which may specify for any particular situation in speech that any of several outcomes may occur depending either on fixed probabilities or on unknown outside factors.

Those rules we have indicated in Appendix H. VI, therefore, are not those for any particular speaker, but rather those of many different speakers with many different characteristics. As such they represent the first recent attempt to present the euphonic combination of natural speech in an orderly fashion that might be understood by a general speech recognizer. Such a presentation, based only on available data, cannot be considered definitive or complete. In some cases there is

not enough information about rules of euphonic combination previously developed to represent them exactly, and certainly a considerable amount of additional research is needed in this area. The rules compiled in Appendix H. VI are thus important particularly for indicating the feasibility of such ordering of acoustic and phonetic data for speech recognition and for outlining potential areas of further exploration.

With the tentative development of such rules, and although we now leave the linguistic aspects of our model to examine the acoustic aspects of analyzing speech for general transcription.

SECTION 4: ACOUSTIC CONSIDERATIONS OF SPEECH

INTRODUCTION

Our object in this section is to relate the acoustic information about speech to physical means of articulation and the rules of euphonic combination. Such correlation, while involving both a review of past research and utilization of concepts developed according to the dimensions of physical articulation, is oriented primarily toward the future development of comprehensive theory and an ordered set of data to meet the analytical needs of a general purpose speech recognizer. Although by no means covering all aspects of acoustic research, therefore, we consider those lines of acoustic investigation that may be most relevant to the generation of further data for transcription of carefully articulated or continuous speech, while suggesting potentially useful avenues of further exploration.

Since the development of vocoders at the beginning of World War II, there has been much interest in identifying speech by its characteristic energy patterns as recorded on spectrograms; investigation in this field has led to a considerable amount of data relevant to our model as well as the development of several limited purpose speech recognizers discussed below. In many aspects of speech production and perception, however, present concepts need further clarification to fit the needs of adequate general recognition by electronic equipment.

One particularly pressing question is the relation of past acoustic analysis to the needs of our multi-dimensional model. Published research in acoustics tends to concentrate on spectrographic analysis of discrete words, some nonsense syllables, and only a few selected sentences. Generation of comprehensive information on speech production and perception must utilize the acoustic data obtained from discrete words, but it must also depend on the concepts of dimensions and continuous physical interaction of sounds which occur in continuous speech. It should be noted that data based on concepts of continuous speech can have a direct relevance to the acoustics of discrete words when such words have more than one syllable; the same transition of voiceless stop to voiced stop occurs in sit down and cupboard.

Integration of past studies with present acoustic work can thus define more clearly the informational needs of speech research in several related fields, notably methods of articulation, methods of perception, and the needs of a speech actuated machine. The results of our integration of such aspects of speech suggest three general needs: (1) Extension of acoustic studies using carefully articulated joined words as well as discrete words; (2) Modification of present sets of

phone classes used for actuating machinery through speech and for evaluating speech processing equipment: (3) Continued investigation into physical manner of articulation.

While the first need cited above involves primarily an orderly increase in the available data by recording speech on spectrograms and time-amplitude plots, the second and third call for a considerable amount of complex research that can relate aspects of present acoustic research to the actual physical production of sound. Present analysis suggests the probable value of reducing the amount of data necessary for speech recognition by organizing acoustic information according to concepts designed in our model. A further efficiency in general transcription of speech is made possible by anticipating the effects of combination of words or syllables through stored acoustic data.

Our research has examined the properties of phone classes such as dental n which have undergone little acoustic analysis in English because they do not distinguish between the meanings of words, and phone classes such as visarga vowels which are not generally recognized by linguists, acousticians, and phoneticians. Although non-distinctive, such phone classes are identifiable both by a distinct set of criteria for physical articulation and by a recognizable speech wave-form; dental n differs from alveolar n both in place of articulation and in the second formant transition. Since a general model for speech recognition must use distinctions acoustically more precise than those of phonemics, the acoustic correlates of such phone classes are important in the automatic transcription of sound.

While acoustic information now available cannot give a comprehensive description of English speech in terms as complete as those described above, present information still needs to be ordered in terms capable of accounting for the detailed acoustic information mentioned in the above paragraph. Without such categorization, involving a reduction in choices necessary to identify a sound, the almost astronomical varieties of phone classes which may need to be identified could transcend the capacities of computing equipment that now exists or that can be projected in the near future. To account for such problems we have already projected an orderly arrangement of phone classes in our model; this section presents a portion of the available acoustic data to justify such an ordering.

In clarifying our method for arranging phone classes according to their physical means of production, we face the third need already cited -- further investigation into physical manner of articulation. The function of the tongue in forming mouth cavities is particularly significant

in the acoustic production of distinctive patterns of energy concentration at certain levels of frequency and certain instances of time. The precise role of this function, however, is not yet fully defined in relation to its modification of the speech wave patterns. The significant contribution of MIT and the Royal Institute of Technology in this field are not to be discounted: we suggest merely an extension of their methods and a modification of some aspects of their work to accomodate the new concepts which have been introduced through our model.

I. BACKGROUND OF ACOUSTIC WORK

Although many of the concepts in this field have been under investigation since the start of studies of generation, of propagation, and of perception of sound, recent investigations in acoustic phonetics seem to be closely connected with instruments that are similar to Dudley's vocoder.

The vocoder is essentially a bank of band-pass filters that divide the speech spectrum into about twenty bands. It was developed primarily for secure voice communication with the use of bandwidths which are a small fraction of those used for the conventional telephone system. (The importance of vocoders in acoustic and phonetic research is a subject of Appendix J). Early success with vocoders increased the interest in understanding the basic nature of speech and of its recognition. The first extensive work on this subject, reported in the textbook Visible Speech by Potter, Kopp and Green, used a modification of the vocoder system -- namely a spectrogram -- for presenting spectral densities of speech waves as functions of time. The speech waveforms studied therein were identified by the words or the sentences spoken and by speaker identities.

Human observers were found capable of "reading" these patterns and also of relating these to those portions of the spoken sentences that produced their waveforms. This was found to be true even for speech waves generated by a number of speakers selected for these tests. Such results indicated the possibility of specifying characteristics of these patterns as representative of certain articulatory positions of the tongue, the lips, the mouth cavity, and of vocal flap vibrations and nasal resonances. Considering the expected and observed variations in speech waveforms of different speakers saying the same sentence, it is to be expected that only the gross characteristics of these patterns were used in studying their relationship to the above aspects of speech production and to speech perception. Information obtained from such studies has

also influenced the work of linguists and phoneticians in their classification of sounds of different languages. One such work is that on the distinctive features of "phonemes."

Some machines that could be actuated by speech were also designed from information about characteristic patterns of speech waves. A brief summary of such activity is presented in Figure 18. The possibility of designing machines that could be actuated by speech opened a new field. (The subject is discussed further in Appendix K).

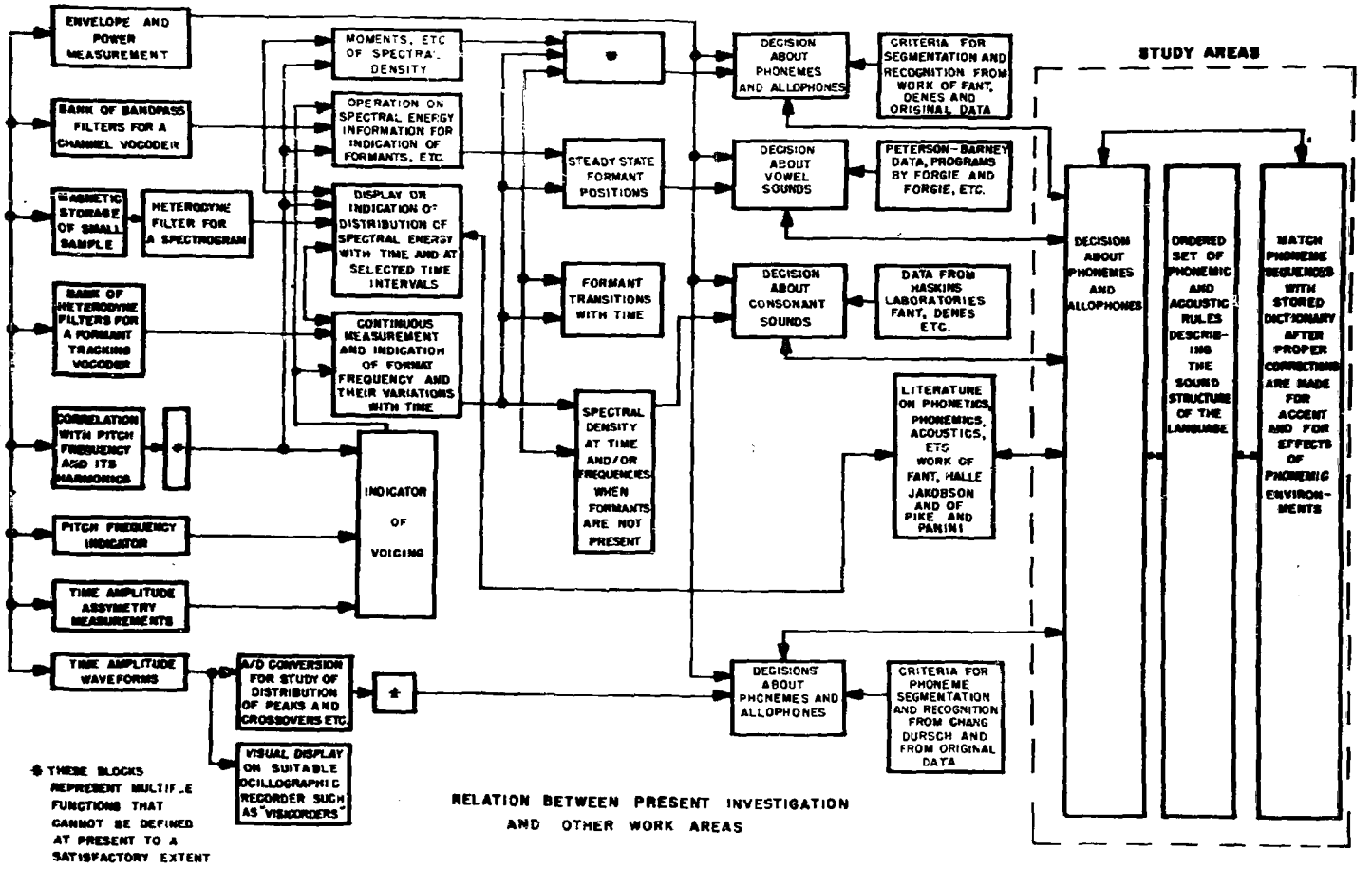
Speech recognizers have been built which assume that a machine is capable of recognizing words or phonemes. Most of these machines have enjoyed limited success, as we mention in Appendix K. That is, most of these machines work with a limited vocabulary and/or with carefully articulated, isolated words. It has often been thought that refining the existing methods would increase the applicability of these machines. However, a more useful approach might be to design a method of computer operation to account for anticipated acoustic imprecisions of speech.

II. COARTICULATION AND DURATION

The present extent of acoustic data on speech production suggests the value of an organized analysis to correlate what is available, what is important, and what is still needed for the development of a general purpose recognizer. The following discussion suggests such an approach; Appendix L relates the information obtained from such work to needs outlined in Section I. At present it may be noted our sources of acoustic information are primarily spectrograms. The particular value of spectrographic analysis is that it presents wave-forms in a pictorial representation which enhances some of the more significant acoustic characteristics of speech. These characteristics have been related to speech perception by work at Bell Telephone Laboratories and Haskins Laboratories, and related to speech production by work at Bell Telephone, Massachusetts Institute of Technology, and the Royal Institute of Technology, Stockholm.

Potential limitations of the spectrograms, however, are suggested by the fact that in the process of presenting information that is most significant to the efficient reception of speech we must also decide upon information that is to be considered redundant. Moreover, in passing speech through a bank of filters used by the spectrograph, speech characteristics are distorted.

Although such decisions and distortions may eliminate information that seems not to be essential for human perception of speech, such as intensity, duration, and rate of articulation, this information may be particularly significant in the interpretation and identification of speech wave-forms by a general purpose transcriber. Thus it should be kept


ENVELOPE AND POWER MEASUREMENT
BANK OF BANDPASS FILTERS FOR A CHANNEL VOCODER
MAGNETIC STORAGE OF SMALL SAMPLE
HETERODYNE FILTER FOR A SPECTROGRAM
BANK OF HETERODYNE FILTERS FOR A FORMANT TRACKING VOCODER
CORRELATION WITH PITCH FREQUENCY AND ITS HARMONICS
PITCH FREQUENCY INDICATOR
TIME AMPLITUDE ASSYMETRY MEASUREMENTS
TIME AMPLITUDE WAVEFORMS
MOMENTS, ETC OF SPECTRAL DENSITY
OPERATION ON SPECTRAL ENERGY INFORMATION FOR INDICATION OF FORMANTS, ETC.
DISPLAY OR INDICATION OF DISTRIBUTION OF SPECTRAL ENERGY WITH TIME AND AT SELECTED TIME INTERVALS
CONTINUOUS MEASUREMENT AND INDICATION OF FORMAT FREQUENCY AND THEIR VARIATIONS WITH TIME
INDICATOR OF VOICING
A/D CONVERSION FOR STUDY OF DISTRIBUTION OF PEAKS AND CROSSOVERS ETC.
VISUAL DISPLAY ON SUITABLE OCILLOGRAPHIC RECORDER SUCH AS "VISICORDERS"
*
STEADY STATE FORMANT POSITIONS
FORMANT TRANSITIONS WITH TIME
SPECTRAL DENSITY AT TIME AND/OR FREQUENCIES WHEN FORMANTS ARE NOT PRESENT
±
DECISION ABOUT PHONEMES AND ALLOPHONES
DECISION ABOUT VOWEL SOUNDS
DECISION ABOUT CONSONANT SOUNDS
DECISIONS ABOUT PHONEMES AND ALLOPHONES
CRITERIA FOR SEGMENTATION AND RECOGNITION FROM WORK OF FANT, DENES AND ORISINAL DATA
PETERSON-BARNEY DATA, PROGRAMS BY FORGIE AND FORGIE, ETC.
DATA FROM HASKINS LABORATORIES FANT, DENES ETC.
LITERATURE ON PHONETICS, PHONEMICS, ACOUSTICS, ETC WORK OF FANT, HALLE JAKOBSON AND OF PIKE AND PANINI
CRITERIA FOR PHONEME SEGMENTATION AND RECOGNITION FROM CHANG DURSCH AND FROM ORIGINAL DATA
STUDY AREAS
DECISION ABOUT PHONEMES AND ALLOPHONES
ORDERED SET OF PHONEMIC AND ACOUSTIC RULES DESCRIBING THE SOUND STRUCTURE OF THE LANGUAGE
MATCH PHONEME SEQUENCES WITH STORED DICTIONARY AFTER PROPER CORRECTIONS ARE MADE FOR ACCENT AND FOR EFFECTS OF PHONEMIC ENVIRONMENTS
* THESE BLOCKS REPRESENT MULTIPLE FUNCTIONS THAT CANNOT BE DEFINED AT PRESENT TO A SATISFACTORY EXTENT
RELATION BETWEEN PRESENT INVESTIGATION AND OTHER WORK AREAS

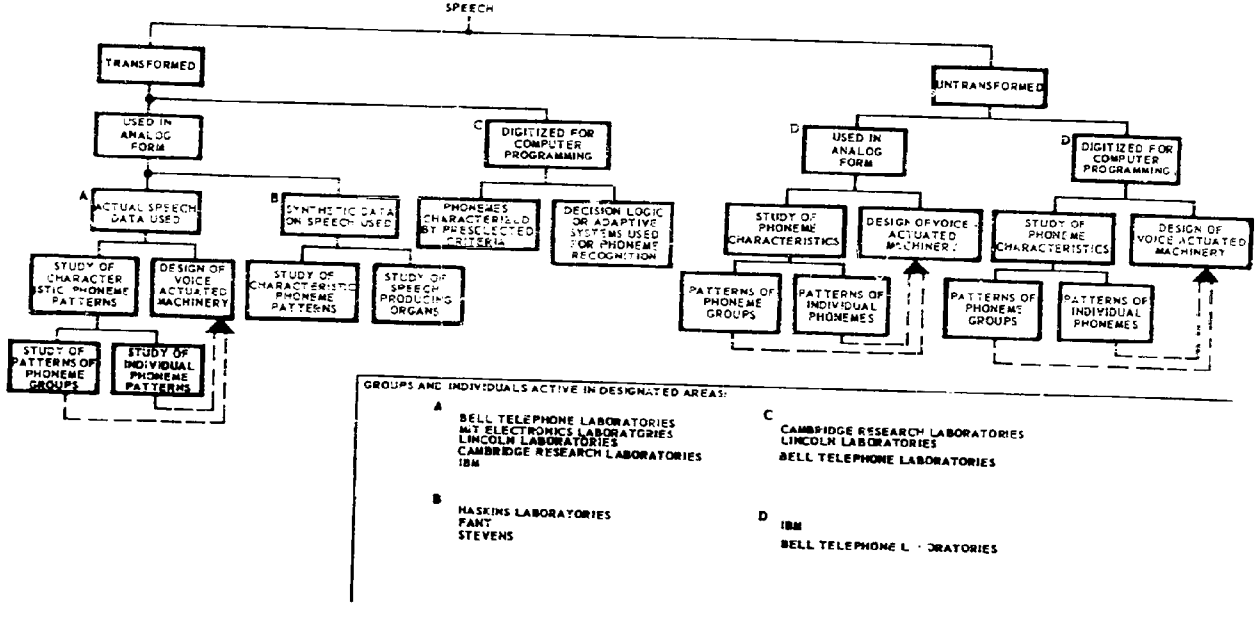


Figure 18b Recognizer Development

in mind that the data presented below represents only one type of acoustic measurement of speech. (Another potential source of information is provided by time-amplitude plots, which represent graphically the frequency of the whole speech wave measured against time.)

On the basis of available acoustic data with the limitations noted above, the following research suggests coarticulation as a necessary description of real speech events and indicates the importance of duration as a dimension of our model. The substantiation of coarticulation from such data is particularly worth noting, since almost all the speech measured was in the form of short words or carefully articulated sentences "manufactured" to illustrate certain phonetic sounds. Natural speech, it should be remembered, tends to run adjacent phones together considerably more than the data discussed below; phonetic transcription of such speech will be discussed in this section.

A. STUDIES IN THE IMPORTANCE OF COARTICULATION

Our model is organized on the assumption that vowels and consonants may be articulated simultaneously and at times from the same physical position in the mouth, as illustrated by the Multidimensional Model. Initially there were two reasons for such an ordering of speech. The first is common observation that many consonants -- particularly labials such as p - allow the tongue to take the position of a following vowel during consonant articulation; a subsidiary support for coarticulation is provided by related experience with Sanskrit phonetic rules of word combination. The purpose of research discussed in this subsection is to provide further evidence for such coarticulation.

If consonants and vowels can be coarticulated, it should follow that there is both physical and acoustic evidence for such coarticulation. That is, not only would we be able to show evidence of coarticulation through palatograms and motion picture X-rays of articulatory motions, but such physical coarticulation should have a definite relation to the acoustic signal of the speech-wave as measured by spectrographs, formant vocoders, or time-amplitude plots. Formant positions of consonants, for example, might be expected to show certain variation in their transitions depending on the characteristic formant positions of following vowels; time-amplitude plots might indicate a characteristic sound wave for each vowel-consonant combination. Although complete investigation of these phenomena is not available, a survey of present data does suggest the presence of such evidence for coarticulation.

Present physical evidence for coarticulation is provided by the work of H. M. Truby, (Truby, 1959) who carried out acoustic and phonetic investigation to determine whether all phones are influenced by adjacent phones. In one such experiment which he describes in detail Truby took motion picture X-rays of articulation of the word plotch ; these X-rays revealed that at the time the lips burst open for p the tongue is already in position for the following l. Similar evidence is available for other consonant clusters in which the physical articulation of one consonant does not interfere with assuming the articulatory position of the following consonant.

Truby's proof of coarticulation exists only for consonants and semi-vowels, but it has long been assumed by phoneticians that speakers tend to reduce the time and energy expended in physical articulation. Thus in circumstances when there is coarticulation of a consonant plus a following consonant or semi-vowel, there is also likely to be coarticulation of a consonant plus following vowel, assuming this is feasible under the physical conditions of articulation. Since the position of the tongue is independent of the articulation of labials, for example, we may expect coarticulation in the word patch as well as in plotch . Such physical coarticulation is also likely to occur for consonants other than labials, though probably to a lesser degree.

Evidence from many sources, moreover, suggests the possibility of correlating physical coarticulation discussed above with specific features of the acoustic signal as measured by spectrographs, other spectral measuring processes, and time-amplitude plots. Although no exact correlation linking physical coarticulation and differences in the levels of transitionary formants presently exists, there is partial acoustic justification for further work on this matter in the research of Ilse Lehiste and Gordon Peterson (Lehiste and Peterson, 1960), Bjorn Lindblom (Lindblom, 1963), and in our own correlation of data from Truby and Visible Speech .

The work of Ilse Lehiste and Gordon Peterson, specifically discussed below, was undertaken to investigate whether an acoustic distinction between formant movements existed to serve as cues for consonant identification and formant movements which signal the presence of a complex syllable nucleus such as a glide or a dipthong. Actual experiments depended primarily on the articulation of one subject, who pronounced specified words within the "frame" sentence: "Say the word _____ again." Spectrograms of these sentences were then made and the researchers measured levels of the first three formants at the following points: start of the onglide measured at the

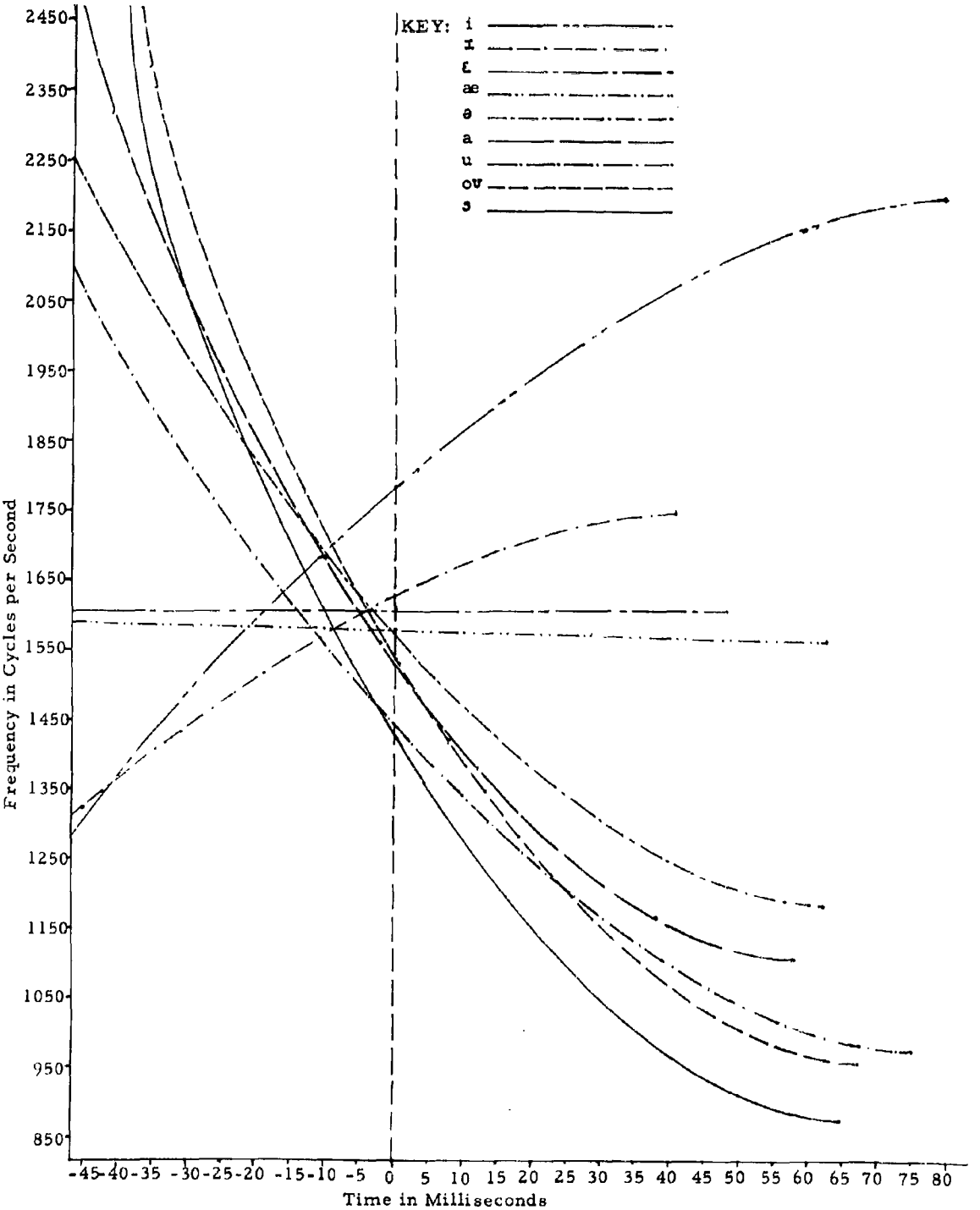


Figure 19 Application of the Locus Theory to Natural Speech

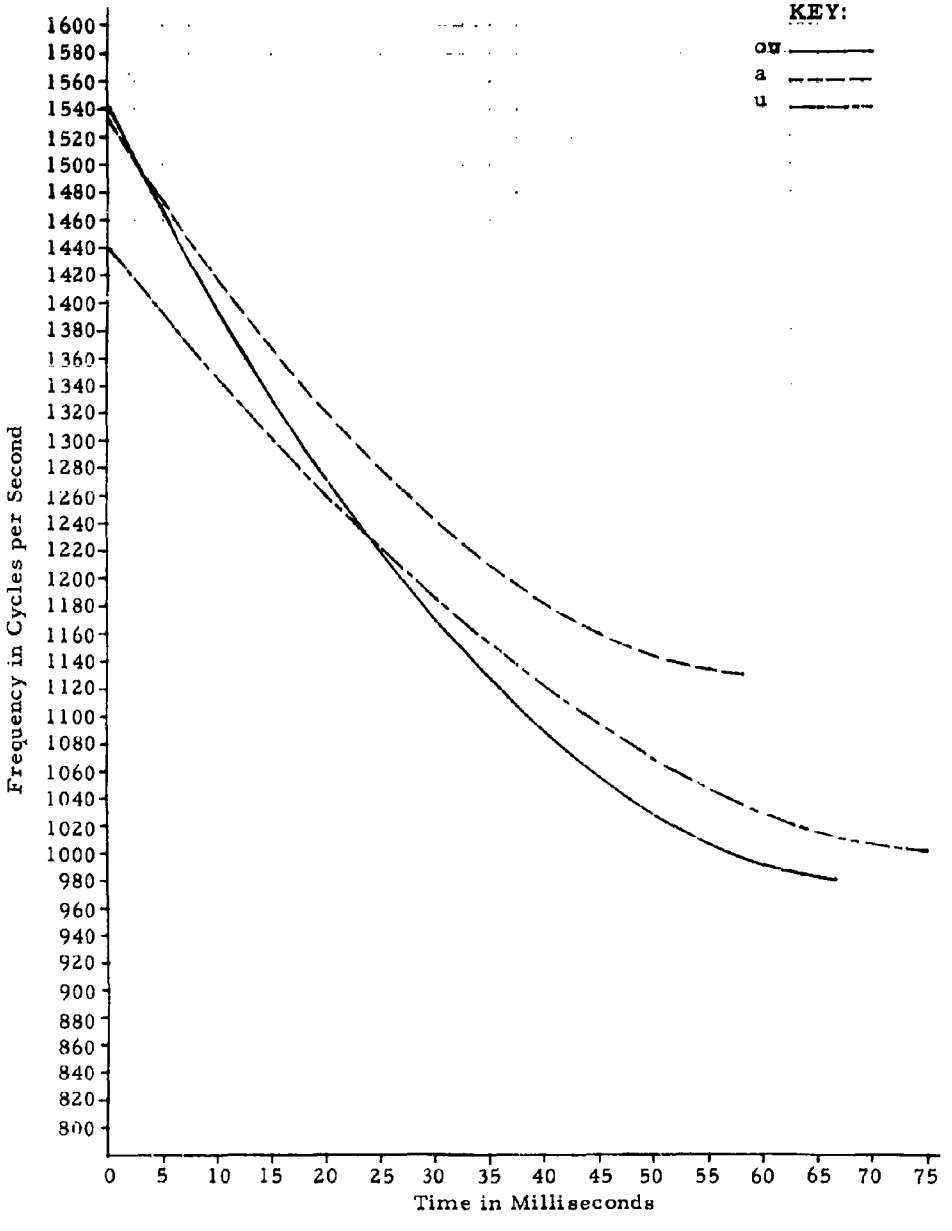


Figure 20 Application of the Locus Theory to Three Vowels in Natural Speech

consonant release; end of the offglide; duration of the onglide from consonant release to steady-state of the syllable nucleus; duration of the steady-state; formant positions at the steady-state; duration of the offglide.

The results of such experiments, shown on the graphs of Figures 19 and 20, indicate the various starting points of the onglide and the steady-state, as well as the duration for different vowels after the consonant d. Lehiste and Peterson discovered that there is a wider frequency range of starting positions for formants of vowels following labials than for vowels following other consonants. and also that the average duration of onglides is shorter. This is most conveniently explained by assuming physical coarticulation of vowels and consonants, since the tongue is not used in articulating labials, it is freer to assume different physical positions than while pronouncing consonants articulated primarily by the tongue; such opportunity for coarticulation has greater effect on formant transitions.

Lehiste and Peterson's work, it may be noted, dealt with individual words articulated, if not discretely, at least in a sentence position set off by pauses. In further investigations into the nature of formant variation experiments with continuous speech are needed. Such research (which also tends to suggest the coarticulation of vowels and consonants) was performed by Bjorn Lindblom at the Royal Institute of Technology, Stockholm.

Lindblom's work was carried out to test the hypothesis that the articulation of vowels in unstressed syllables is centralized -- that is, occurs at the mid-point of the traditional phonetic chart for vowel articulation. He made spectrographic analyses of one subject pronouncing consonant-vowel-consonant combinations such as kak under varying timing conditions and with a systematically varying context. Having tabulated his spectrograms, Lindblom was able to compare the formant onglides of the vowels of such minimal pairs as bob and gog; his results indicated considerable frequency variation in the steady-state of identical vowels positioned between different consonants.

It is probable that the formant levels of Lindblom's vowels were affected by following consonants, but his work also supports our belief that the articulation of consonants and following vowels is interdependent. Representative measurements taken by Lindblom (On Vowel Reduction, p. 35, Figure 11) for the vowel a between b and g; the frequency level for the vowel steady-state of gag is approximately 350 cycles higher than that of bab when the vowel duration is very short. Such measurements, duplicated in Lindblom's work with other vowels, also show that the influence of a preceding consonant on a vowel steady-state will have an inverse relation to the duration of the vowel.

In order to ascertain whether there may be a difference in the acoustics of given consonants when articulated with different vowels,

we made measurements of various formant levels of consonant-vowel combinations as indicated on spectrograms from Truby, Visible Speech, and Lehiste and Peterson. A discussion of our method for performing these measurements may be found in Appendix M. Although a certain factor of error in measurement must be allowed, and although the data examined is by no means exhaustive, we noted a considerable variation in the frequency level at the beginning of the onglide of various vowels following the same initial consonant. For example, we examined the frequency level at the starting point of the voiced second formant onglides following the consonant b : for these starting points there is a variation which ranges from 640 cycles for the onglide of ə to 1925 cycles for the onglide of i. We further noted energy concentrations at varying frequencies within the voiceless portion of s when it occurs in combination with various vowels and semi-vowels. Such concentrations occur well below the greater energy levels of the unvoiced portion of s, which gather at regions above 2500 cycles; the frequency range of the lower energy concentrations can vary from between 230 cycles for s of sweet to 1750 cycles for s of see. Such variation suggests the influence of vowels on preceding consonants. It also suggests the existence of a second formant onglide for combinations of voiceless consonants plus vowels.

Tabulation of the data presented above tends to offer a body of evidence in conflict with the locus theory (Delattre, Liberman and Cooper, 1955) which maintains that regardless of what vowel may follow, all formant transitions from any specific consonant begin at one specific frequency level for each consonant, although the initial part of this onglide cannot be measured. According to this theory, for example, all the formant transitions following the consonant d have a characteristic slope which, if extended backward, would meet at one point called the locus, (Delattre, et. al., "Acoustic Loci and Transitional Cues for Consonants," J. Acoust. Soc. Amer., P. 771, Figure 4.)

Such a theory if valid would make it feasible to identify preceding consonants by the slope of the onglide of the following phone; if vowels and consonants are coarticulated, however, the frequency level of onglides would seem to depend upon the physical articulation of the following vowel, rather than characteristic onglide slope pointing to the locus. Since the resolution of such a conflict is important to the identification of acoustic aspects of speech, we discuss it in further detail below.

The locus theory was developed by Haskins Laboratories as part of a method for stylizing formant patterns which could be put onto the Pattern Playback to produce sound waves which listeners heard as specific speech sounds. A primary purpose of these experiments was to define those characteristics of the speech wave important in the perception

of phone classes; by actuating speech through synthesized formant patterns Haskins was able for the first time to conceptualize and define the importance of significant elements of the speech wave as start of the onglide, frequency at start of the onglide, deviation from start of the onglide to steady-state, and location of the center frequency of unvoiced portions of the noise burst.

Within the confines of this important research the locus theory represents a system for ordering data about the acoustic patterns of the speech wave in terms suitable for experimentation. As it is not feasible to construct stylized formant patterns without imposing some form of order on their location, the locus theory was eventually conceived as a successful effort to organize acoustic patterns of the sound wave from the Playback to be identified by listeners. Since the experiments of Haskins were all with stylized speech, however, the question remains whether the results obtained from such perceptual studies can directly describe the characteristics of actual speech.

In applying the results of Haskins' Pattern Playback to general speech one would need to rely on two assumptions. The first is that the locus to which the transitionary formants lead will not be affected by the following vowel. The second premise deals with the nature of the stylized formant patterns which produced sounds identified by listeners as falling within given phone groups; one might need to assume that such stylized patterns produced through mechanical methods can be used to give information about the acoustic characteristics of speech as it is produced by human beings.

At present there is insufficient data to prove or disprove the second premise. The first premise is called into question by the data on physical coarticulation shown by Truby and also by the variations in formant levels which were found by Lehiste and other researchers in combinations of one initial consonant with different vowels. The experiments of Lehiste and Peterson also suggest that in all cases the formant transitions of given consonants will not, if extended back, meet at one locus. In Figure 19 (which we have constructed according to the results published by Lehiste and Peterson) shows the different transitionary onglides from the consonant d; the onglides have been extended but they do not meet at one point. Moreover, in plotting the formant onglides of the vowels ou, a, and u according to data from Lehiste in Figure 20, it is apparent that the onglides cross each other in that portion of the speech wave which is actually measureable.

Further data to be considered are the experiments at Haskins itself with Pattern Playback (Delattre, Liberman, and Cooper, 1955). The experimenters synthesized spectrograms of speech. On these stylized spectrograms the first formant onglide and steady-state was kept constant, while the second formant was drawn at various levels.

of frequency. In all cases transitions to various vowels were drawn, originating from the supposed locus of d rather than simply pointing to it.

When these patterns (with complete formant transitions) were run through the Playback, they produced sounds identified by listeners as b, d, g, and d, depending on the relative frequency level of the formant steady-state for the following vowel. When the initial portion of the transition was erased, however, all sounds were heard as combinations of d and following vowels.

Haskins explains this data by suggesting that part of the change in the position of the articulators to go from a consonant to a vowel takes place during a silent period before the measureable beginning of the on-glide. Thus transitionary formants only point back to a locus rather than leading there. Such evidence, however, might also indicate that the starting point of a formant transition after consonants depends on the articulation of following vowels.

Our particular criterion for using coarticulation as an acoustic division of our model was availability of physical and acoustic data tending to indicate evidence for such coarticulation. In the review presented above, there would seem to be present sufficient data to justify continued use of coarticulation. Having thus reviewed the acoustic and phonetic aspects of coarticulation we will explore in the following section the effects of a phone's duration on the acoustic characteristics of its wave form.

B. THE ACOUSTIC EFFECT OF CHANGES IN DURATION

Available evidence suggests that duration measurements may be particularly important in distinguishing between different vowel-consonant combinations, compensating for individual differences in stress, and identifying words uttered by speakers using different dialects. The need for making such important distinctions in transcribing speech supports our inclusion of duration as a dimension of our model.

The relation of duration to consonant-vowel combinations has been particularly explored by Björn Lindblom who recorded consonant-vowel-consonant syllables under various conditions of stress, intensity, and duration of a vowel and the level of the final formant position reached in its articulation depend on the preceding and following consonants. According to Lindblom, moreover, the relation between the steady-state level of the second formant of a vowel and its duration may be described by a mathmatical expression derived from curve-fitting techniques. Such relationships indicate the important acoustic nature of duration in speech production.

Lindblom's work, moreover, indicates that the duration of on-glides, offglides, and steady-state are likewise affected by preceding and following consonants, but the effect of such influence is also related to the tempo, mood, and emphasis of a particular speaker. Such data raises complex problems in constructing a general transcriber, since it will be necessary to analyze the enunciations of individuals who will have different rhythms of speech and different dialects.

Work by Peterson and Lehiste (previously mentioned and also discussed in Appendix N) has provided additional information on duration based on spectrographic analysis of consonant-vowel-consonant combinations articulated by one speaker in a pre-selected environment. Such research provides an approach to automatic phone classification. Lehiste has suggested that the ratio between the steady-state, and off-glide of any particular vowel will remain constant. The Peterson - Lehiste data, however, is limited to articulation of a small number of isolated words by one speaker in an extremely controlled environment. Additional work is needed to apply such data to a general purpose recognizer, whose problems in analyzing various dialects and speech rhythms have been noted above.

Spectrograms of a standard British, General American, and Southern pronunciation of the same utterance, shown in Section 2, Figure 9, for example, indicate that the British speaker has comparatively abrupt formant transitions and a long steady-state, while the Southern speaker has long transitions and almost no steady-state. These data seem to indicate that Lehiste and Peterson's conclusions about General American are not necessarily applicable to other dialects.

Another factor which makes the measurement of duration important is that in a consonant cluster a consonant may be shortened. We have evidence, which is presented in Figure 21, that this shortening affects l, r, and w after voiced stops, and it may affect other consonants also. However, the duration of the vowel following the cluster is apparently not affected.

One additional aspect of duration which again seems to suggest its importance as a dimension of our model is the possibility of comparing the duration of one vowel to the duration of another as one criterion for distinguishing between them. Acoustic measurement contained in Appendix N indicate that the duration of r is always longer than the duration of l in the same environment.

III. THE APPLICATION OF AVAILABLE ACOUSTIC DATA TO THE NEEDS OF A GENERAL PURPOSE RECOGNIZER

Our interest in studying the correlation of auditory phonetics and acoustics is to extend present knowledge about the characteristics of sound combination, particularly in terms of data useful to a general model for speech recognition. Such an extension involves two steps - the orderly classification of past experimentation as it relates to such problems as slurs, segmentation, sound drop-outs, or sound change, and the extension or modification of such experimentation to include a more precise definition of the varieties of phones likely to appear when two or more sounds occur in combination.

Considerable work has already been published on the acoustic patterns of words (generally of one or two syllables) uttered "discretely" - one at a time. Such work helps to define the initial acoustic characteristics of specified words and in some cases contributes to the construction of machines for limited transcription such as the digit recognizers.

As has been suggested above, however, the articulation of a particular phone class can be influenced by preceding or following phones; the words sit down may be heard as sidown when spoken as a phrase. Such modification of sound may also cause the acoustic pattern of discrete words to differ from those of words or phrases in which two or more syllables occur together, as in the case of cupboard, which exhibits the same transition of voiceless stop to voiced stop that has been observed in sit down.

Since the construction of a general model for speech recognition requires information about the wave-forms of phone classes as they occur in combination, there is an apparent value in data relating to the acoustics of carefully articulated speech and speech as it occurs in normal conversation. We have used the data of the work of Potter, Kopp, and Green in Visible Speech for our information about the acoustic correlates of euphonic combination. (Published results of the investigation in Visible Speech takes the form of wide-band spectrograms representing careful enunciation of sentences; in their work it was actually possible to train observers to "read" sounds represented by these spectrographic patterns whether as a whole sentence or as fragments of sentences).

Before discussing the relevance of Visible Speech research to our model certain aspects of its basic data should be noted. Speech which Potter, Kopp, and Green transcribed consists of phrases either carefully selected or specially prepared to illustrate certain phonetic combinations. The sentence "Have half above five," for example, was

composed simply to illustrate acoustic differences between h, f, and v. The words in Visible Speech, moreover, are articulated with considerably more precision than in normal speech; in the sentence When did you cut the wheat ? there is a measurable pause of at least 85 milliseconds between complete decay of the final t in cut and the onset of the, although most speakers would combine the sounds either in careful reading or normal conversation. Speech articulated with such extreme precision may not entirely reflect the acoustics of general speech.

Such speech, nevertheless, does provide a valuable basis for the formulation of data about sound waves and sound substitution. Even with extremely precise articulation as in Visible Speech it is possible to secure tentative acoustic confirmation for a number of the phonetic substitutions which our rules of euphonic combination suggested as likely to occur in continuous speech.

Close examination of the spectrograms from Visible Speech provides evidence for such changes as the voicing of a normally voiceless consonant (indicated by the presence of harmonic energy in the very low frequency ranges) or the substitution of a stop for a spirant (indicated by the absence of a stop gap). In several instances such actual changes in carefully articulated continuous speech reflect phonetic changes, based on physical means of articulation, that we have already suggested as probable. For example, we suggest the rule that "a voiced consonant can become voiceless if it occurs before a voiceless stop, spirant, or sibilant;" our example of a situation in which this might happen during continuous speech is big town . An examination of spectrograms from Visible Speech gives the acoustic indication of such a substitution in the sentence, This is such a big church (p. 156); in this actual example there is no voice bar for the "s" of is , so that the voiced consonant has become voiceless and the word is , normally pronounced "iz", may be perceived as "iss." Other instances where available acoustics information supports our suggested rules of sound change are discussed in Appendix O of this report.

At present, it should be noted, most of our data is limited to reduced illustrations of spectrograms from Visible Speech; such illustrations give more information about aspects of speech production indicated by random energy or discontinuities than about those aspects dependent on formant transitions. Thus the present confirmation of rules of euphonic combination is confined mainly to manner of articulation and resonances, which produce such readily identifiable acoustic characteristics as full closure of sound, energy bursts, frictional energy, or characteristic placement of formants. Information about

place of articulation requires the generation of additional acoustic data, particularly in relation to spectrographic analysis. Time-amplitude plots, seldom analyzed in relation to continuous speech, may also yield helpful information, possibly about place of articulation and almost certainly about manner of articulation and resonance.

The limitations of data on the acoustics of carefully articulated continuous speech and the relative success in confirming rules of euphonic combination with such limited data suggest the need for further research in this area. One additional impetus to such research is the possibility that analysis of continuous speech may reveal or emphasize the importance of acoustically distinct phone classes not generally recognized by phoneticians. Research discussed in Appendix O has already supported the existence of visarga and stressed the distinct acoustic characteristics of dental n̲. Information about such phone classes may be helpful both in the accurate transcription of continuous speech and in the identification of regional variations likely to occur in discrete words.

Although such a discussion can only summarize the most important aspects of the relation between acoustics and phonetics, the value of further experiments to extend data from discrete words is apparent. Such research might involve investigation into the acoustic patterns of an orderly set of coarticulated phone classes in polysyllabic words. With sufficient information of this type at our disposal it could then become feasible to evaluate such data as characteristic acoustic patterns as they occur in continuous speech. Such research can enable us both to derive further examples of euphonic combination as it occurs in natural speech and to apply this data to the acoustic equipment of a general purpose transcriber. Some aspects of research recommended for immediate work in this area form the basis of the following discussion.

IV. APPLICATION OF THE RULES OF EUPHONIC COMBINATION TO CONTINUOUS SPEECH

The final tests of rules of euphonic combination are whether these rules describe situations that commonly occur in general speech, and whether such rules may be effectively utilized for the analysis of speech by a general purpose speech recognizer. Research discussed in Appendix P is accordingly designed to test euphonic rules previously reported by applying them to both carefully articulated and conversational speech. The method of such application is analogous to the analytical processing of speech by a general purpose recognizer and suggests the comparative advantage of using euphonic rules rather than word-unit recognizers of speech.

In our research recordings of carefully articulated speech and of normal conversational speech were studied; the respective texts were "What is a Boy?" and "What is a Girl?" read on a 45 rpm record by Jackie Gleason, and experimental tape recordings of conversation made by Dr. J. M. Pickett in an anechoic chamber at Hanscom Air Force Base, Bedford, Massachusetts. Phonetic transcriptions by ear of selected portions of this speech were then made by a phonetician to discover examples of euphonic combination previously suggested. A summary phonetic analysis of the carefully articulated and the conversational speech yielded examples of at least twelve different forms of euphonic combination previously suggested. Among these examples (cited in Appendix P) were significant confirmation for changes in resonance, changes in place of articulation, and the rejoining of the final phone in one word to the initial phone of the word following.

Such auditory analysis, it should be noted, is only a preliminary test to suggest the value of previous research to speech recognition in terms of our model. Additional rules may be confirmed by subjecting the speech waves to careful acoustic instrumental analysis. By ear alone, for example, it is difficult to identify the existence of full stop closures, or place of articulation, as illustrated in the differences between dental and alveolar n. Further data will rely on acoustic measurement as well as phonetic transcription.

The Multidimensional Model for organizing speech information represents a departure from Western methods of describing speech and constructing speech recognizers. We reject the assumption that it is possible to construct a phoneme recognizer or one which can recognize words as separate units consisting of phonemes. Evidence indicates that adjacent phones influence each others' articulation; thus it does not seem feasible that received sounds be transmitted directly to a dictionary of stored acoustic information. The preliminary transcription of speech must first be subjected to criteria for sound change and euphonic combination, criteria whose relation to the dimensions of place and manner of articulation, intensity, duration, and resonance, has been mentioned earlier in this report. Such initial processing provides an efficient and necessary method for solving problems such as those caused by sound changes within polysyllabic words with the phones of adjacent words.

Essentially the operations of such a recognizer as that described above may be segmented into four main steps (1) symbolic representation of rules received; (2) application of rules of euphonic combination to separate words slurred together during pronunciation; (3) processing these units through an electronic "dictionary" to identify their meaning; and (4) written transcription of speech. Research discussed in Section 3

indicates the feasibility of representing the sounds of speech production symbolically. The phonetic analysis of continuous and carefully articulated speech cited above indicates that such transcription using orderly rules of euphonic combination may be necessary for an accurate rendition of speech as it is generally produced. Such an assertion is reinforced by the need to analyze speech in detail for identification of slurred sounds, by the proven ability of rules previously developed to suggest euphonic combination occuring in randomly chosen texts, and by the need for rectifying errors in transcription or identification through an intermediary that can apply acoustic data to speech transcription and provide alternatives to combinations of sounds that the "dictionary" cannot identify.

Such a concept, it will be noted, is in conflict with theories of speech recognition that rely on identification of word units alone. These theories assume it may be possible to achieve an effective recognizer by constructing a dictionary to contain the stored acoustic pattern of words most commonly used in English up to ten thousand words. Although such a recognizer might be used to identify the extremely careful articulations of a few highly trained individuals, its application to general speech must necessarily be complicated by euphonic combinations and individual varieties of pronunciation. These problems are discussed below.

In research with continuous and carefully articulated speech by linguists and phoneticians it has been a general observation that euphonic combination and coarticulation are natural phenomena of speech; a corrolary of this observation is that unstressed words are usually combined with adjacent words. In accordance with these criteria it may be noted that such words as to, it's or is are incorporated into the articulation of surrounding words in such phrases as t'learn, 's too hot, or 's thata fact?

While it is possible to assume that such small unstressed words can be articulated one at a time, some analysis and practice will satisy the reader that such articulation is unnatural and difficult even for a trained speaker. Word-unit recognizers, although expensive to construct, could thus operate only under special conditions and could not be applied to many speech situations. An additional limitation to the potential value of word-unit recognizers is suggested by the fact that the articulation of a particular phone depends on many variable factors based on the physical conditions of articulation. Such factors include intensity, duration, resonance, and place and manner of articulation, whose importance has already been emphasized in their organization

as dimensions of our model; because of their effect on speech production it is to be expected that no two speakers will pronounce the same word in exactly the same fashion; it is probable that individual articulations will often be sufficiently distinct that they produce a variety of acoustic patterns not readily related.

Substantiation for the effect of such dimensions cited above is indicated by coarticulation, varied emphasis, and euphonic combination discovered through phonetic transcription of conversation and carefully articulated speech. From the careful speech of one speaker, for example, it was possible to record six phonetically distinct pronunciations of the word of within a one-minute interval, particularly as it occurred in the phrase of every. Such variations depended particularly on word placement within a sentence, stress, and the rhythm of the speaker in voicing his ideas. Phonetic transcription also yielded examples of a euphonic combination within words, as in the loss of h in grasshopper, the transformation of a voiced h to a voiceless p in absolutely -- both examples drawn from our recording of careful speech -- or the substitution of a glottal stop for a t before a labial in the word voltmeter in the Hanscom recording of conversational speech.

From the data discussed above several observations may be suggested. The first is the difficulty of applying techniques used in word-unit recognizers to the transcription of general speech. Evidence for such difficulty may be found particularly in the number of slurs and euphonic combinations both between words and within polysyllabic words.

A second observation relates to the problems inherent in articulating discrete speech for transcription by a word-unit recognizer. Even in careful speech there may be many instances of euphonic combinations between words; and it may also be particularly difficult while articulating speech to be transcribed by a unit recognizer to avoid the inevitable euphonic combinations that take place within words, as in the loss of t in softness from the Gleason recording.

Additional problems for word-unit recognizers are also indicated in the facts that one speaker may pronounce the same combination of words in several different ways, as in the phrase of every from Gleason, while two speakers will almost certainly pronounce various identical words in ways which are phonetically and acoustically distinct; phonetic transcription indicates this to be the case with the word cars pronounced by two different speakers in the Hanscom recordings (Passage1).

If we were to assume arbitrarily that factors discussed above caused only eight possible modifications of stored acoustic patterns for given words, it is apparent that word-unit recognizers would need to contain a considerable amount of redundant data. By organizing such data according to divisions which our model has been using, however, we intend to improve the efficiency in speech transcription while applying our knowledge of euphonic combination to a larger syllabus of words than that available with unit recognizers. Experiments discussed above, it may be noted, simply provide an initial outline of how our multidimensional analysis is substantiated in its application to practical problems of word recognition.

Several necessary lines of further investigation are evident. Among these are complete compilation of a symbolic representation of speech production, generation of additional rules of euphonic combination, and generation of our own data describing the relations between acoustic and phonetic aspects of speech. These steps will form the basis for our continued investigation.

In the available evidence, there seems to be considerable justification for a unified approach to speech analysis, based on the genetic, phonetic, linguistic, and acoustic aspects of speech. In order to obtain acoustic and phonetic substantiation for treating the articulation of phones as complex phenomena described by an orderly set of rules based on various physical means of production, additional data must be examined. In this subsection we present and discuss information we have generated concerning the acoustic correlates of phone classes.

This data, which had previously been undefined or even unidentified, is essential to the conceptual completeness of our model for speech recognition. With our increased knowledge of these phone classes, we are better able to categorize them. The accuracy with which speech segments can be identified is increased and the number of choices required to identify a sound is reduced. Moreover, such information suggests that minor adjustments can be made in certain rules, adjustments which would increase efficiency -- both by making certain rules more widely applicable and by refining other rules to apply to special circumstances. Certainly this new empirical information describes only a limited number of phone classes and is not yet complete enough - in a statistical sense - to make adjustments obligatory in certain cases. But this information nevertheless broaches areas that we would like to study for making additional refinement. Although the small amount of data, which is limited to demonstrating peculiarities of speech events, constrains our conclusions, we do present a small sampling of rule

occurrence. This represents the first information we have been able to gather on the possibility of rules for more effecient computer use. This possibility must be investigated, by obtaining additional information about the probability of their being operative, in circumstances where they had previously been mentioned to operate.

Tha fact that we could correlate our analysis of acoustic modification as represented on spectrograms and time amplitude plots with our perception of how and when sound change actually occurred made the generation of our own data extremely worthwhile. The need for such correlation has been indicated a number of times when discussing our use of source data. This work is an extension of acoustic studies using joined words. Thus, one result of our present data analysis is the demonstrable empirical proof it provides for the deductive reasoning - based on source data - by which we evolved our rules of euphonic combination and our concept of coarticulation. This enables us to proceed with even greater assurance in the construction of our model. Our data analysis also confirms that the concept of coarticulation is essential in describing speech production.

Whereas our concern in euphonic combination is with such broad problems as the elision of a sound or the general fusion of sounds, our concern in dealing with coarticulation includes the minute influences of one sound on another in its environment. Our descriptions of such minute influences in the following discussion are intended to verify coarticulation in general: but it is necessary to extend such isolated contributions into a system of rules -- through continued research on the characteristics of coarticulation, By generating more object-directed data such as this, we may be able to encompass details with rules, ultimately reducing the number of rules - as we have done with euphonic combination - to the point where a computer can store and apply them. We envision a time when we will be able to increase the efficiency of speech transcription by anticipating all coarticulation through stored rules of coarticulation. The development of such rules however, would be a study in itself.

In addition, coarticulation as an approach to segmentation is of great importance: our division of will you into the coarticulatory segments [wi][llyou] is an example of segmentation by coarticulated entities, a new approach to segmentation. So the principle of coarticulation will be useful to a machine not only in the reception but in the analysis and segmentation of sounds in the speech flow.

As a result of our analysis of the characteristics of various speech segments -- and the precise representation which would make them

suitable for computer programming -- we are able to incorporate the phone classes h, r, l, and w into our model. At the start of our work, we were not certain of how these segments should be positioned. Thus in Section 1 we deferred a discussion of these "problem segments" Our work with precise representations has helped us in defining their positions in the model.

Finally we will discuss methods by which our rules of euphonic combination might be employed in a working computer system. (Of course our rules will also be used in programming for the computer.) Once the best method, with accompanying protection, verification, and efficiency techniques has been decided, the novelty and usefulness of such a program -- even outside the scope of developing a general purpose speech recognizer -- cannot be minimized. However, the methods must be selected with care. At the present we are certain only of alternative possible methods, each of which has its deficiencies and its advantages. We present them in summary form in this report, while we continue to work out more detailed problems each presents. To develop a working computer system is beyond the scope of the present study, but it is a subject that merits investigation.

A. DISCUSSION OF SPEECH DATA

When we use data from other people's literature, a good amount of time is required in tracking down, culling and reapplying this data to fit our particular needs. Of course, we are occasionally confronted with instances in which the data we desire is either not adequate or not available. Even when we are able to gather sufficient data from other research efforts, we are for the sake of accuracy forced to ascertain the precision of measurements given. And in a number of instances, the published work in measurement is incomplete for our purposes. For example, no one, to our knowledge, publishes information derived from the use of time amplitude plots: although we have found them very valuable, their usefulness appears to have been overlooked or ignored by others. Furthermore, other people's data provides no information about the effects of intensity on the coarticulation of phone classes - for no one reports intensity in a form we can use: such information is vital to us since we wish to know what part intensity has in distinguishing phone classes: we have already stated that duration may differentiate sounds (such as the vowels of bomb and balm in certain American dialects) but we must know the effect of intensity on duration before we can develop a recognition plan as sensitive as the one our aims require. Finally: few people have recognized coarticulation: the only available studies of coarticulation deal with extremely isolated cases: so we are forced to execute our own tests in order to

study acoustical data for the speech phenomenon central to our work. In fact, whenever possible, data generated for any experiment was subjected to analysis for the study of coarticulation.

B. DATA ON "PROBLEM" PHONE CLASSES

In Section 1 we deferred our discussion of certain problem segments. There we mentioned the greatly variant h and r, the elusive complex semivowels y, w, ŋ, m, n, l. Work on our computer program has led us to include h and r with the semivowels; like the semivowels, among other reasons for this, h and r, can only occur before or after vowel sounds (barring isolated exception such as h w in where, phonetically transcribed hwere.) When writing our first report, we were not certain about the classification of the semivowels as a whole: should they be treated as consonants? or as a separate manner of articulation? Now it seems likely that we will treat them as a part of the vowel cluster in which they occur. For example, the word crush would be segmented cru sh . Of course we are faced with many vexing questions beyond the matter of workable classification. For instance there is the problem of the doubtful existence of semivowels in particular environments. We cannot assume they exist because of our orthographic tradition. Consider characteristics of "y". There is little acoustical doubt that absolute initial y exists. But in some environmental circumstances - particularly when it is by or between [i]sounds - there is no immediate clearcut evidence of its true presence. If it does exist, how is it represented acoustically? Are there sound waveform manifestations that machines can identify? Another problem is that the nature or the occurrence of a semivowel may vary with speaker or dialect. Can we develop rules to predict these variations? and to account for them? First we must have a more definite idea of the acoustic and articulatory properties of each semivowel. This is an immense task beyond the scope of our present study. In our initial report we felt that we would be able to learn more about h, r, ŋ than we actually were able to learn. On the other hand, we felt then that we actually would not be able to treat vocalic or "syllabic" nasals at all - and further on in this section we present data on the vocalic nasals. Below we discuss our work on the w and y phone classes.

In our work with the w and y phone classes, we focused our attention on those environments in which a w or a y might or might not exist. Evidence of their existence in such environments would be good evidence of the basic acoustic characteristics of these phone classes. And the latter is our ultimate concern.

1. The w Phone Class

For this test, a phrase containing w was compared with a phrase not containing w but otherwise identical. The particular phrases choses were "no ax" and "no wax." These phrases were particularly well suited for our study because the frequency level of F_2 (second formant) at the end of the vowel of 'no' is very close to the F_2 level of w. In fact, the similarity between this vowel and w is not only acoustical, but also articulatory; for both require rounding of the lips for speech production. Therefore, if present here, w is forced to distinguish itself, to make itself known.

The selected phrases were incorporated in the sentences "we have no ax" and "we have no wax." The sentence containing "no ax" was included in a list of sentences which five informants read at the beginning of the recording session. After that, the same informants read a list of three or more sentences, including the "no wax" one. There was an interim of at least twenty minutes between the reading of "no ax" and "no wax." This was to deter the possibility that the informants exaggerate differences between the phrases. (See Figures 21-26 for w and Figures 27-30 for y. See also Table 1 for a detailed description of the duration changes for "no ax - no wax" for each speaker and Table 2 for a close measurement of "no ax - no wax" frequency changes. Comparable for y study are Tables 3 and 4.)

After making tape recordings of these readings in studio, we made spectrograms of the sentences that concerned us on a Kay Sonagraph. Spectrograms were made for the speech of all five speakers, but those made for Speakers 3 and 4 were not measured; these informants were women with high-pitched voices and we found it difficult to make formant measurements accurate enough to be at all conclusive.

Differences between these two phrases were similar in enunciation by each of the three speakers. In all cases "no wax" has a steady-state in which the second formant is at a very low frequency and has very weak intensity (see Figures 21, 23, & 25). The duration of the steady-state ranges - with different speakers - from 65 to 106 milliseconds. Speaker 2's pronunciation of "no ax" has a second formant steady-state of similar frequency and intensity; the duration of this steady-state is 32 milliseconds. Speakers 1 and 5 pronounced "no ax" with a steady-state in which the second formant is at a slightly higher frequency and has much greater intensity (see Figures 22, 24, & 26). The duration of this steady-state

by Speaker 2

Figure 24

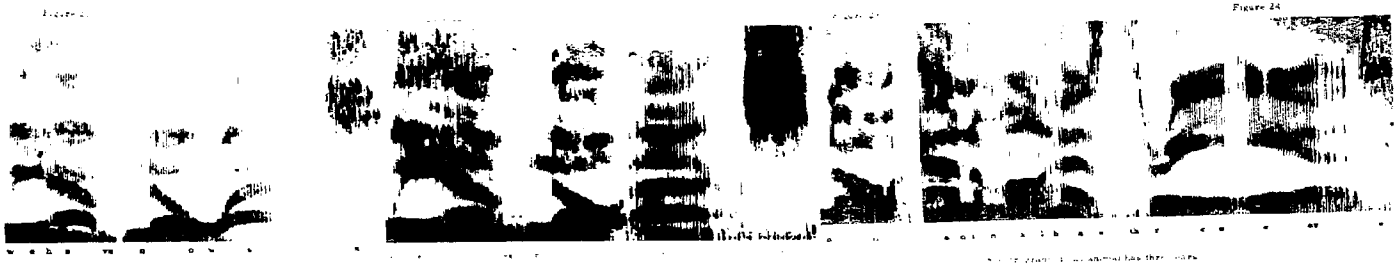


Table 1. Comparison of Duration Changes for "no ax" - "no wax"

	A. Duration in milliseconds for:				
	F_1	Steady-State	Offglide	F_2 Steady-State at Low Frequency and Very Weak Intensity (All Other Formants Very Weak Intensity)	F_2 Steady-State at Low Frequency with Normal Intensity
Studied Data:					
Speaker I					
"no ax"	84	42	97		65
"no wax"	40	42	107	26	
Speaker II					
"no ax"	45	26	104	52	
"no wax"	47	18	98	54	
Speaker V					
"no ax"	64	48	154		58
"no wax"	64	48	107	106	

	B. Duration in Milliseconds for:				
	Onglide before Pause	Pause (All Formants Weak or Absent)	Onglide	Steady-State	Offglide
Studied Data:					
Speaker I					
"no ax"	14	66	55	100	65
"no wax"			100	23	26
Speaker II					
"no ax"			87	81	81
"no wax"			77	71	29
Speaker V					
"no ax"	16	71	32	129	77
"no wax"			100	42	Not Measurable

Table 2: Comparison of frequency changes for "no-ax" - "no wax"

	A. Frequency in Cycles per Second at:					
	Start of Steady-State	Start of Offglide	End of Steady-State	End of Steady-State	Before Pause	After Pause
Studied Data:						
Speaker I						
"no ax"	1742	1742	1606	1606	1606	2044
"no wax"	1636	1635	606	606		
Speaker II						
"no ax"	1431	1431	648	648	648	676
"no wax"	1376	1437	746			
Speaker V						
"no ax"	1643	1643	641	641	1224	1531
"no wax"	1593	1643	746			

Table 2: Comparison of frequency changes for "no-ax" - "no wax"

	B. Frequency in Cycles Per Second at:		
	End of Onglide	End of Steady-State	End of Offglide
Studied Data:			
Speaker I			
"no ax"	1721	1721	1686
"no wax"	1795	1795	1686
Speaker II			
"no ax"	1655	1655	1686
"no wax"	1624	1624	Not Measured
Speaker V			
"no ax"	1600	1624	1655
"no wax"	1593	1624	Not Measured

Table 3. Comparison of Duration Changes for "three ears - three years"

A. Duration in Milliseconds for:

Studied Data:	n	r	Onglide	Steady-State (Normal Intensity)	Fr. Steady-State at Start of Frequency and Low Intensity	Pause (All F changes were Weak or Not Present)
Speaker I						
three ears	74	52	77	120		64
three years	94	42	67	94		
Speaker V						
three ears	74	26	74	126		
three years	81	21	74	19	81+	

Table 3. Comparison of Duration Changes for "three ears - three years"

B. Duration in Milliseconds for:

	Steady-State (Normal Intensity)	Glide	Steady-State
Speaker I			
three ears	36	87	
three years	101	109	84
Speaker V			
three ears		90	84
three years		132	48

Table 4. Comparison of Frequency Changes for "three ears" - "three years"

A. Frequency in Cycles Per Second at:

Studied Data:	Start of r	Start of Onglide	Start of Steady-State	Steady-State Level before Pause	Steady-State Level after Pause
Speaker I					
"three ears"	1407	1500	2160	2250	2250
"three years"	1170	1500	2250	2280	2280
Speaker V					
three ears	1240	1310	1000		
"three years"	1275	1372	1938		

Table 4. Comparison of Frequency Changes for "three ears" - "three years"

B. Frequency in Cycles Per Second at:

Studied Data:	Steady-State Level Before "fadeout"	Steady-State Level After "fadeout"	End of Steady-State	End of Offglide	End of r Steady-State
Speaker I					
"three ears"			2250	1686	
"three years"			2280	1715	1686
Speaker V					
"three ears"			2155	1710	Not Measurable
"three years"	2031	2248	2248	1924	1841

ranges from 23 to 26 milliseconds. For both Speaker 1 and Speaker 5 the vowel onglide that follows the "no ax" steady-state is interrupted by a pause (a period during which all formants are greatly reduced in intensity). This pause lasts 65 milliseconds for one speaker and 70 milliseconds for the other.

So the differences between "no wax" and "no ax" are summarized as follows: for Speaker 2 there is a significant duration difference between the two steady-states; for Speakers 1 and 5 the apparent differences are in the duration, the frequency level, and the intensity of the second formant steady-state as well as the pause in the following onglide. All these differences identify w. We still need to refine this identification by gathering more information - particularly about differences in transition (See Appendix Q).

2. The y Phone Class

The method employed for investigation of y was the same used in the study of w. In fact, the two experiments were done with the same informants at the same recording session. In our study of y we used the phrases "three ears" and "three years" as incorporated in the sentences "No animal has three ears" and "It lasted three years." It was not until some time after the experiment that Speaker 2's spectrograms were found to be imperfect. And our analysis of the speech of Speakers 1 and 5 (the only two remaining in the y solution attempt) put us no closer to an understanding of the distinctions that verify y's presence than we were at the outset. Examination of the spectrograms for these two speakers revealed no consistent differences between the phrase with y and the phrase without it. On the other hand, the spectrograms of the two phrases show a major similarity. For both speakers there is only one vowel steady-state for the entire phrase; although for Speaker 1 this steady-state is interrupted by a pause in "three ears." (This steady-state is not entirely level; it shows a slight rise in all cases.) See Figure 27.(but see Appendix L, §12).

What if a difference between the two phrases does not exist? What if y disappears in this environment leaving us with homonyms? How do we teach the machine to solve the homonym problem? It would have to ignore information that may be pertinent elsewhere. Frankley, we would prefer to eliminate homonyms -- to prove that in every case there are significant identifiable differences, but we may have to recognize that in some instances it will be impossible to eliminate such homonyms.

3. Syllabic Nasals

We have a spectrogram and a time-amplitude plot of a syllabic n in the word kitten spoken by Speaker 1 (See Figures 31 and 32). The spectro-

gram shows that this n has a strong F_1 at 300 CPS and another fairly strong formant (probably F_5) at around 5000 CPS. All the formants in between are extremely weak. Other n's by the same speaker have more energy between 300 and 5000 (see Figures 21, 22, 27). The time-amplitude plot of the syllabic n shows a wave-form which is very different from that of an ordinary n spoken by any of our informants for this research (see Figure 33 for an example by comparison). We have a time-amplitude plot from earlier work which shows an n with similar syllabic waveform in the phrase "moon" (See Figure 33). This n is followed by an n with an ordinary waveform.

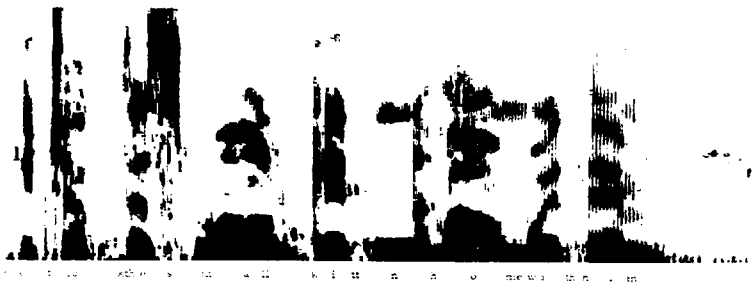
Incidentally the reader will notice the brief vowel-like portion immediately following the syllabic nasal on our spectrogram. This portion may be the result of releasing the oral closure before the celum is closed.

4. A New Vocalic Portion

The matter of this unclassified speech-sound is so problematical that a linguist transcribing speech generally overlooks or ignores it. But a speech transcribing machine could not ignore it unless instructed. Not only does this vocalic portion show up on a spectrogram, but it has very distinctive characteristics on a time-amplitude plot (see Figures 31 & 32). Furthermore, it appears frequently so in building a speech recognizer we must plan our data to allow for its occurrences.

A tentative explanation of the existence of this speech sound is the sudden closure of the velum while sounding of a nasal is not complete. Take the word kitten for example. There is little change in tongue position for the t and the n, both being aveolar. At the end of the t sound's friction the velum opens all the way; and this results in the nasal resonance we identify as n (Nasality occurs when more air flows through the nasal passage than flows through the mouth cavity.) But the velum closes before it was to close for the end of the nasal; there is an accidental flow of air through the mouth cavity creating a new phone class. Such a phone class may have spectral characteristics that are similar to those of a nasalized vowel articulated in a similar manner, as discussed by Fant. This situation occurs when the mouth cavity impedence is comparable to that of the nasal cavity coupled to the oral passage by a limited opening of the velum.

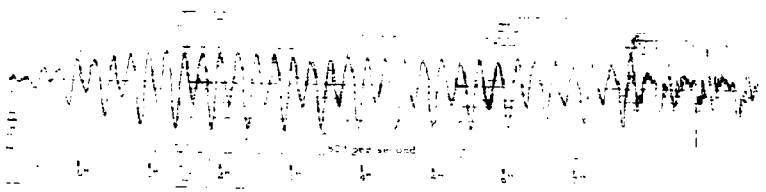
Since the vocalic portion lasts only from 3-4 pitch periods, it is difficult to define, with reliability, its spectral, formant characteristics (see Figures 31 & 32) or other aspects of its waveform. Yet it is difficult



Spectrogram of "He took the small kitten home with him"

of Speaker I

Figure 3



Time-amplitude plot of syllabic n of "kitten" in "He took the small kitten home with him"
of Speaker L

Figure 4

40" per second

Time-Amplitude Plot of Syllabic n from
Articulation of "moon"

Figure 53

to instruct a machine to ignore this short-duration vowel. Three or four pitch periods is the duration of the (I) vowel in the word animal (see Figures 27 & 29). If the machine ignores the vowel-like portion, it will ignore the vowel (I) in the word animal. We can construct a valid workable rule only if we give the machine more information about the vocalic portion than its duration alone; information that will show how this nonsense segment is different from cognitive segments. Tentatively, our rule in this instance would state that after a nasal any vowel-like segment of five pitch periods or less must be ignored unless the segment following that one is a nasal (animal); or unless it is sandwiched between two voiceless plosives (like pit); or between a voiceless plosive and a nasal (pin); or vice versa (nip).

The occurrence of phenomena like this vocalic portion help to emphasize the problems of people at work on phoneme recognizers. They can teach their equipment to ignore such phenomena -- but it is necessary first to understand and to classify the phenomena beforehand.

5. Classification of Problem Segments

After having examined the acoustic data on certain problem segments, we are better equipped to undertake their classification in the multi-dimensional model. In Section I we listed h, r, y, w, and vocal m, n, ŋ and l, as consonants which were difficult to fit into the model (we are calling these consonants simply because they occur in those parts of words more often occupied by consonants than by vowels; they were not classified as consonants on any acoustic basis).

The problem of h was relatively simple to solve because the consonants in the model are grouped with their following vowels, the fact that the place of articulation of h (and consequently the quality of h) changes with every vowel is no longer problematical. We thus classified h as a voiceless vowel - or the voiceless portion of whatever vowel follows it. Examples from Visible Speech (Chapter 9, Unit I, pp. 110-118) show that the frequency of h is the same as the frequency of the steady-state of the following vowel - except that, in most cases, the h is unvoiced. However, between two vowels (as in the sentence "Will you help us?", the h may be voiced. This voiced h is a special category of speech with measurable characteristics and must receive special treatment.

The International Phonetic Alphabet mentions only three vowel-r combinations, all of which are closely related; the ɚ sound in church, the ɝ sound in bird as pronounced by a Southern American or an Englishman and the ɹ̩ in better and similar unstressed syllables.

We believe, however, that for the identification of r (as well as certain vowel sounds) it is important to notice that the vowel sounds in such words as art, glare, fear (or true, tray, trouble) can be diphthongized with r, so that a machine may not easily distinguish where one sound ends and the other begins. The transition from the steady-state of the vowel to the steady-state of the r (or vice versa) seems an important clue to the recognition both of the vowel and of r. We therefore treat r as a portion of the vowel (i. e. a vowel cluster). However, an additional Riemann leaf should be included in the model to indicate the retroflex manner of articulation.

In the acoustic representations of the sentences "No animal has three ears," and "It lasted three years," presented above, we tried to determine whether y was a semi-vowel pronounced as a diphthong with the following vowel. The time-amplitude plots and spectrograms, as we have seen, showed that no y can be conclusively distinguished: "Years" and "ears" seem to be acoustic homonyms in this context.

In the sentences "We have no ax" and "We have no wax" a slight break was noticeable in the second formant of the "no...ax" segment of the first sentence; no such pause was usually present in the "no... wax" portion of the second sentence. Thus it seems that w will require special rules for resolution (and possibly will require the use of probabilty). However, it seems that it can be recognized.

We also advise treating l as a vowel cluster; again a separate Reimann leaf has been included in the model to specify the lateral manner of articulation. Evidence of frequency and duration measurements to be mentioned later in this section substantiate this treatment.

It must be pointed out, however, that in careful speech one can reliably segment an l, such as Gunnar Fant has done ("Studies of Minimal Sound Units"), because the duration and the frequency mark it as a separate entity in that case. But in continuous speech l often last no longer than three or four pitch periods and shows no appreciable change in frequency. From a genetive point of view, one could explain this by noting that in continuous speech, because of the lateral manner of articulation for l, the tongue does not bend enough to yield a significant difference in the acoustic representation. In continuous speech, then, l should be treated as a semi-vowel; in careful speech, it can justifiably be treated as a consonant.

Finally, vocalic m, n and ŋ will also be treated as vowel clusters, as described in the case of r. We have not yet worked out the particulars

of vowel recognition, as such an endeavor lies beyond the scope of the present study. However, we have several general suggestions concerning vowels.

(6) Recognition of Vowels

Recent measurements of the frequency characteristics of vowels have indicated that it is difficult to distinguish acoustic areas which correspond to the traditional phonetic vowel classes. The variation in frequency, which results in "overlapping" of closely related vowel classes, seems to be the result of changes in the environment, the rate of articulation, the intensity, and the duration of certain speech sounds. For example, our measurements of spectrograms in Visible Speech showed that the I sound as in bit ranges from 1517 to 2041 cycles per second in the F_2 steady-state, as the consonant environment changes. This evidence does not, however, contradict the vowel recognition program developed by Forgie and Forgie, since their program specifies a limited context and a fixed environment, which would stabilize the frequency of the F_2 steady-state for a given vowel.

In normal speech, however, vowel sounds occur in many environments, with various degrees of stress, which alters the rate of articulation, the intensity, and the duration characteristics. For this reason, it seems advisable to allow for an F_2 variation; this can be done by "broadening" the range for the vowel classes (and hence reducing the number of vowel classes the machine recognizes). But although the "overlapping" of classes would be greatly reduced, a contextual program would need to be formulated to select the correct vowel. Such a program is being developed for consonants; by means of this program unallowable consonant combinations will be eliminated. To develop a contextual program for vowels, however, is beyond the scope of this study.

C. VERIFICATION OF COARTICULATION AND EUPHONIC COMBINATION

(1) "Will"

In one test of coarticulation we studied the word "will" repeated by the same speaker, but in different environments: (1) as an item isolated on a word list; (2) as the initial word in the isolated sentence, "Will you help us?" and (3) as a word in the middle of a sentence in a continuous passage (The specific context was, "We hope, therefore, a judicious reader will give himself some pains to observe...")

Table 5 shows duration and F_2 frequency measurements for the word "will" uttered in these three different contexts. (Figures

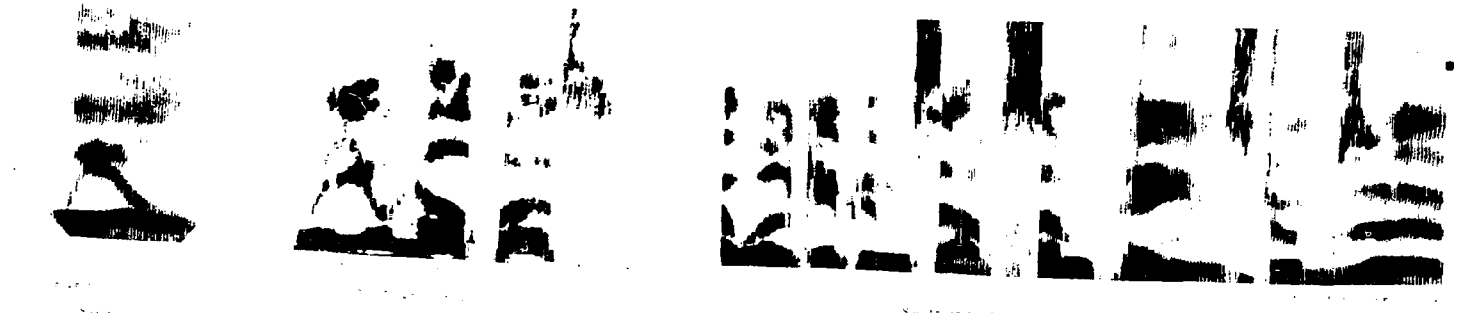
34, 35, 36, 37, 38, 39, 40, 41, and 42) These are measurements for three different speakers. Comparing the acoustic data representing variance in the pronunciation of "will" we note generally that the frequency of the vowel steady-state is highest for the single word, somewhat lower for the sentence, and considerably lower for continuous speech. According to Lindblom, we should expect formant levels to be influenced by duration: and in fact in most cases we can correlate the lowering of formant levels with the decreased duration environmental exigency has imp osed. However, although both the vowel offglide and the l are considerably shortened in the sentence or in the continuous speech as compared to the isolated word, frequency levels rise in both cases. In the case of the sentence environment the rise is particularly acute. And at the same time, the reader will observe that we found it impossible to segment between the l and y of will you in the sentence spoken by Speaker 1 (see Figure 35). So we conclude that the high F_2 level of l was the effect of coarticulated y. The words will you must be segmented wi and llyou, verifying the coarticulation concept. It is the coarticulated lly that affects the vowel offglide of i.

In the cases of the other two informants, it is possible for a human being to perform a very intricate segmentation of l and y. But this would be extremely difficult for a machine to do with reliability. Therefore it is always preferable to segment wi llyou. In both of these cases (Speakers 2 and 5) the F_2 frequency level of l before y is much higher than the F_2 level of l in the other contexts. In the first section of this report (See also Appendix B) our chart of the laterals showed four phone classes of l with four different places of articulation. In the rules listed in Appendix H, there is a rule to the effect that before y and alveolar lateral becomes a palatal lateral. A palatal lateral, like all other palatal sounds, has a high F_2. The high F_2 of the palatal laterals of will you (in those instances where l-y segmentation is possible) has been predicted by our rules of euphonic combination. In fact, the coarticulation of the l-y of Speaker 1 fits the description of the palatal l. (See Figure 38).

Before proceeding to our next example of coarticulation we wish to make two observations. First; we have already mentioned that we intend to include the semi-vowels y and l with the vowel clusters. So the segmentation wi llyou fits our model. Second; it is worth noting that this data negates Ilse Lehiste's hypothesis that there is a constant ratio for the onglide, steady-state, and offglide of the vowel. For all three speakers the offglide is longer than the steady-state in the case of the single word, and shorter in the case of the continuous passage.

(2) l-g

The phrase from which this sound sequence is taken was "will give himself". The spectrogram (see Figures 36, 39 and 42) reveals that there is a closure following the l. Therefore, there is no coarticulation of l before a stop.



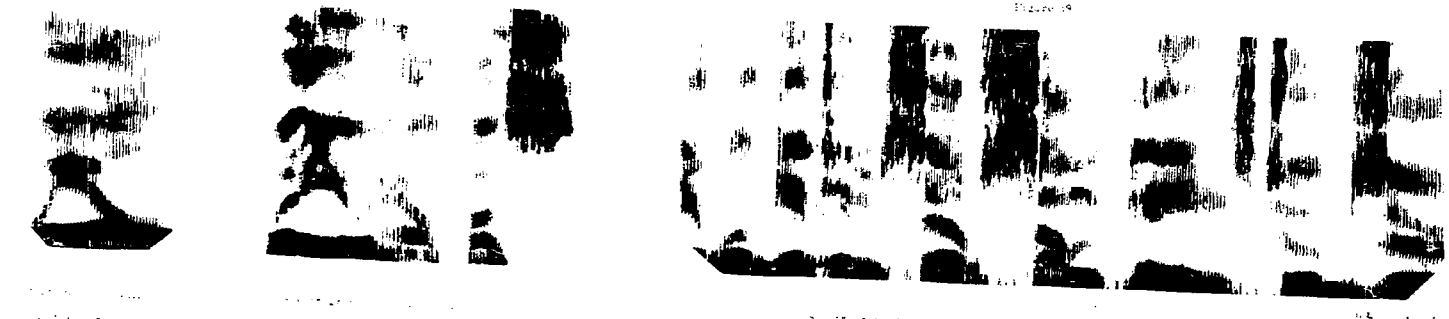


Table 5a: Duration in Milliseconds for the word "will" in three different environments.

Data Studied for "will":	Duration of: w Steady-State	Vowel Onglide	Vowel Steady-State	Vowel Offglide	l
Speaker I					
As a single word	55	42	42	135	217
In a sentence	53		Total [illegible]		+
In continuous speech	45	45	22	39	29
Speaker II					
As a single word	77	[illegible]	[illegible]	122	170
In a sentence	47	45	42	49	85
In continuous speech	41	45	26	54	91
Speaker V					
As a single word	67	54	64	85	170
In a sentence	--	43	22	29	22
In continuous speech	26	26	36	22	16

+ l cannot be separated from following y.

Table 5b: F_2 frequency measurements for the word "will" in three different environments.

Data Studied for "will":	F_2 frequency level at: Beginning of w	Beginning of Onglide	Beginning of Steady-State	Beginning of Offglide	Beginning of l	End of l
Speaker I						
As a single word	--	1196	1809	1809	775	--
In a sentence	--	850	Highest Frequency = 1745		1680	+
In continuous speech	--	711	1292	1324	1292	--
Speaker II						
As a single word	--	678	1780	1780	969	*840
In a sentence	704	970	1421	1583	2067	2067
In continuous speech	--	748	1098	1163	1098	1024
Speaker V						
As a single word	--	740	1780	1840	1030	800
In a sentence	--	680	1580	1650	2100	2070
In continuous speech	710	810	1000	1100	1030	--

+ l cannot be be separated from following y.

*Measured 45 milliseconds before end.

DURATION

	Offglide	w Silent	Onglide	Steady-State	Silent	Silent	Silent	Steady-State	Offglide	Onglide	Steady-St.	Offglid
F_1												
F_2	97.6	45.0	65.0	34.5	9.4	0.27	47.0	0.27	9.4	31.0	47.7	37.6
F_3												

	Steady-State	Offglide	m Nasal	n	Onglide	Steady-State	Offglide	l	y	Onglide	Steady-State	Offglide	Nasal
F_1					37.6	40.7	28.2			21.9	34.5		
F_2	12.5	11.3	87.8	75.2	No	37.6	84.5	47.0	84.5	No	28.2	65.3	45.2
F_3													

Silent	p Burst	Steady-State	Offglide	n Nasal	n	Silent	t	Onglide	Steady-State	Silent	Onglide	Steady-State
F_1											59.6	179.0
F_2	90.1	9.4	40.7	126.6	87.8	72.1	34.5	21.9	37.6	28.2	64.1	47.0	No	234.1
F_3														

Speaker No. 1. will give himself some pains to observe.

Table 5c

FREQUENCY

	Onglide	w Onglide	Steady-State	ll	g	i Steady-State	Offglide	Steady-State	Offglide	Onglide	e Steady-St.	Offglid
F_1										531	717	624
F_2	672	810	1312	1312	1312	2095	1750	1750	1595	No	1624	1218
F_3												

	Onglide	Steady-State	Offglide	p Burst	ai Steady-State	Offglide	t	Onglide	o Steady-State	Onglide	Steady-State
F_1	624	810	775							374	624
F_2	No	1498	1187	1312	2000	2406	2248	1429	1281	No	1593
F_3											

Speaker No. 1. will give himself some pains to observe

Table 5c

DURATION

	Steady-State	Offglide	m Nasal	Silent	s	Steady-State	Offglide	[illegible]	[illegible]	[illegible]	Steady-State	Onglide	Nasal	Silent	p
F_1															
F_2	[illegible]	[illegible]	62.7	[illegible]	[illegible]	31.	[illegible]	25.	47.	122.	[illegible]	[illegible]	78.4	[illegible]	[illegible]
F_3															

	Steady-State	Offglide	Nasal	s	Silent	[illegible]	Onglide	Steady-State	Silent	[illegible]	Steady-State	Onglide	Steady-State
F_1													
F_2	40.7	[illegible]	[illegible]	87.8	[illegible]	124.2	34.5	78.4	[illegible]	[illegible]	[illegible]	62.7	[illegible]
F_3													

Speaker No. 2 himself some pains to observe

Table [illegible]

FREQUENCY

	Steady-State	Offglide	Steady State	Offglide	I	Steady-State	Offglide	ai Steady-State	Offglide	o Onglide	Steady-State
F_1											
F_2	[illegible]	1312	1405	[illegible]	903	1312	1093	2031	2093	1281	1031
F_3											

	Steady-State	er Onglide	Steady-State
F_1			
F_2	[illegible]	[illegible]	[illegible]
F_3			

Speaker No. 2. himself some pains to observe

Table 39

DURATION

	w	Onglide	i Steady-State	Offglide	L
F_1					
F_2	40.7	[illegible]	54.5	147.4	[illegible]
F_3					

FREQUENCY

	w	Onglide	Steady-State	Offglide	L
F_1					
F_2	[illegible]	[illegible]	1906	[illegible]	[illegible]
F_3					

Speaker No. 5 Will

Table 4[illegible]

DURATION

	w Onglide	I Steady-State	U Onglide	y Steady-State	Offglide	ou Steady-State	h Onglide	e Steady-State	l Offglide	Silent	u Offglide	s Unvoiced
F_1												
F_2	[illegible]	[illegible]	43.9	[illegible]	78.4	25.0	40.7	40.7	97.2		72.1	187.8
F_3												

FREQUENCY

	Beginning of Onglide	Steady-State	Onglide	Steady-State	End of Offglide	Steady-State	Beginning of Onglide	Steady-State	End of Offglide	End of Offglide
F_1										
F_2	1062	1250	1843	2062	1218	1124	1374	1750	1000	1342
F_3										

Speaker No. 5. Will you help us.

Table 41

DURATION

	ɛr Offglide	s Silent	w Onglide	l Steady-State	s Silent	ɑ	l Steady-State	l Offglide	ʌ	n Steady-State	Steady-State	i Offglide
F_1												
F_2	[illegible]	[illegible]	26.2	40.7	[illegible]	[illegible]	[illegible]	47.0	[illegible]	[illegible]	22.8	24.2
F_3												

	n Nasal	Silent	s	Steady-State	Offglide	t	ʃ	o (F1) Onglide	Steady-State	Offglide	n Nasal	Silent	p Burst	ai Steady-State
F_1								25.0	31.3	25.0				
F_2	17.1	18.9	116.8	31.3	76.4	97.6	142.4	No	31.3	41.6	87.8	72.1	21.9	166.9
F_3								No	72.1	No				

	n Nasal	s	Silent	t	o Onglide	Steady-State	Sil lent	ʒ	er Onglide	Steady-State
F_1									72.2	109.8
F_2	78.4	72.1	26.2	63.5	63.1	37.1	84.2	97.8	No	179.6
F_3									109.8	72.1

Speaker No. 5. er will give himself some pains to observe.

Tat 42

FREQUENCY

	ɛr Offglide	w Onglide	l Steady-State	l Steady-State	Offglide	h Steady-State	i Steady-State	Offglide	e Steady-State	Offglide
F_1		372	1031							
F_2	1312	—	—	2031	1624	2310	2310	2249	1562	1000
F_3										

	o (F1) Onglide	Steady-State	Offglide	ai Steady-State	o Onglide	Steady-State	er Onglide	Steady-State
F_1	500	717	531				500	531
F_2	No	1312	1031	2499	1405	1187	No	1374
F_3	No	2341	No				2217	1781

Speaker No. 5. er will give himself some pains to observe.

Table 42

(3) m-s

From the same phrase. The spectrogram (see Figures 36, 39, and 42) shows that these sounds are separated; no evidence of influencing each other.

(4) l-f

From the same phrase. Sometimes the l disappears in this environment. In this case (see Figures 36 and 42) we have a back l before the f. The coarticulation (the influence of f upon l) can probably be expected when l is followed by any labial.

(5) o-o

The next phrase analyzed was "to observe". Here the sound sequence o-o (see Figures 36 and 42) becomes a dipthong. We have not constructed rules for the coarticulation of vowel sounds; determination of such rules is beyond the scope of the present study.

(6) b-s

From the same phrase. b and s are almost coarticulated. b's release is weak and short but the moment of release is certainly perceptible (see Figures 36 and 42). s does influence b's frequency level; for here b's energy is near that of the s and of course b usually has its release energy at lower frequencies.

(7) t-l

The phrase studied in this case was "it lasted." t-l is an example of coarticulation in the sense that t is significantly modified by l (See Figures 28 and 30). First; the friction noise for t continues beyond the first voicing pulse of l. Second: t's aspiration is almost absent. Of course the loss of aspiration is frequent when we deal with final t. But this is not a final t; we have shown that its friction noise runs on into l; this t should not be treated as an instance of final t aspiration loss.

(8) s-t

From the same phrase. The rule that reads t following s loses aspiration does not truly apply here. Aspiration is caused by the natural stressing of the phrase as a whole; la is emphasized and sted is not (See Figures 30, 43, and 44). This results in some lack

of definiteness about the e ; there is an occasional aspiration of the t, as here. This observation points out the importance of intensity measurements.

(9) l-h

The sentence considered next was "No animal has three ears," and the effect of h on l studied. But they are definitely distinct (see Figures 27 and 29). h is by necessity initial in English. If l became attached to the next phone class in this case then, h would probably be lost. Since this would be detrimental to comprehension, it never happens, to our knowledge.

(10) s-th

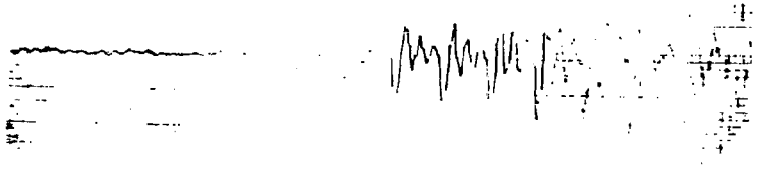
From the same phrase. In the word "has" s is usually a z. Here however the spectrogram (see Figures 27 and 29) shows voicing cessation -- s becoming primarily s. This verifies one of our early rules of euphonic combination, which states that before a voiceless sound a voiced sound may become voiceless.

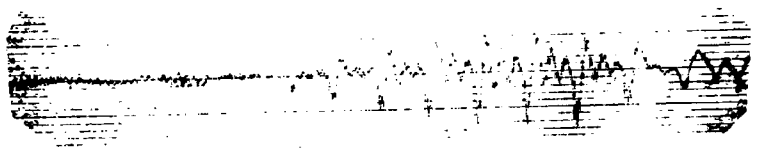
(11) k-th

Finally we concentrated on the sentence, "He took the small kitten home with him." Here the k-th from "took the" is coarticulated. k is very weak here and continues into the th sound: it is extremely difficult if not impossible to segment between the k and the th, either on the spectrogram or on the time-amplitude plots (see Figures 31 and 45).

(12) s-m

From the same sentence. There is a segment of about forty milliseconds before m in which the noise energy of s is extremely decreased or attenuated. This may be caused by the opening of the velum and the consequent side-tracking of principal air flow from the mouth cavity to the nasal cavity. The vocal flap oscillation does not begin until the end of this forty millisecond period. This period might in fact be the phone class linguists call voiceless nasal. But it is very difficult for a machine to classify and to use such a segment of speech. Evidently such a voiceless portion is normally present when s is followed by a nasal -- although the duration of this portion varies from 20-40 milliseconds. The attenuated energy level for the friction of s does not in all cases reach the low level it reaches with the sm of small: this is the influence of m (see Figure 46).

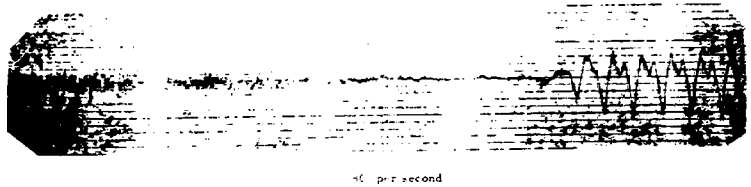




...per second

Time-Amplitude P ot of th_ ot 'it lasted
from "It lasted three years" by Speaker S

Figure 44



...per second

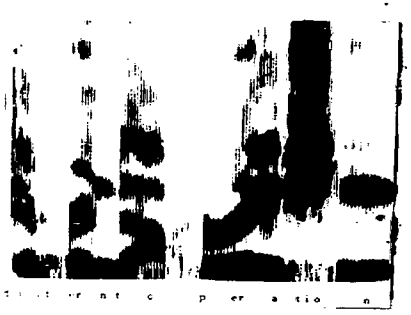
Time-Amplitude Plot of a "voiceless
nasal" between s and n_ in small_

Figure 46

Spectrogram of "...different operations..."

by Speaker 1

Figure 47



Spectrogram of "...different operations..."

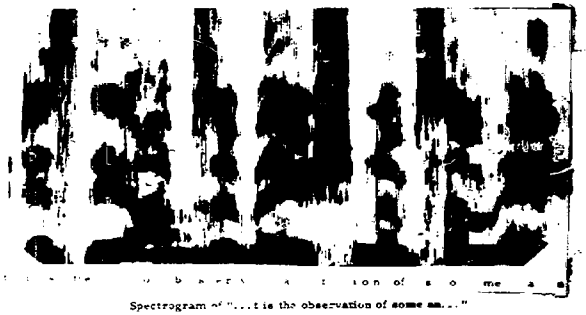
by Speaker 2

Figure 48

Spectrogram of "...various different operations..."

by Speaker 3

Figure 49



Spectrogram of "...it is the observation of some an..."

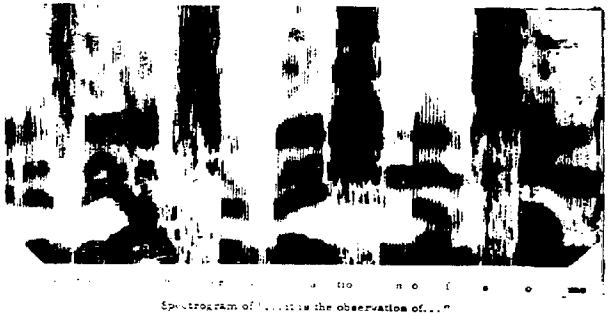
by Speaker 1

Figure 50

Spectrogram of "...various different operations..."

by Speaker 3

Figure 51



Spectrogram of "...it is the observation of..."

by Speaker 5

Figure 52

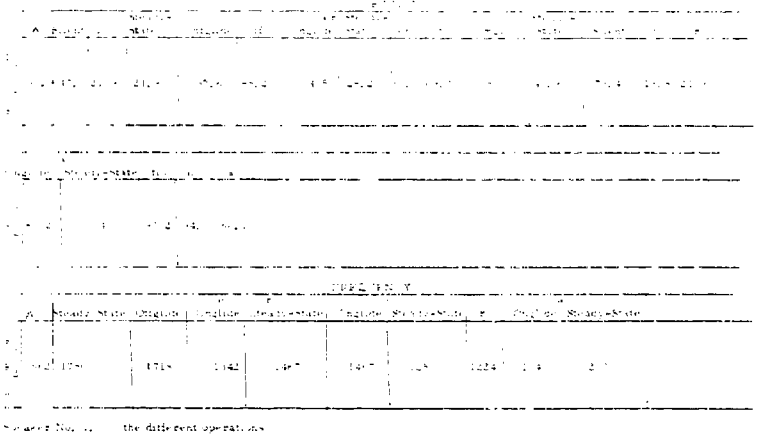


Table 47

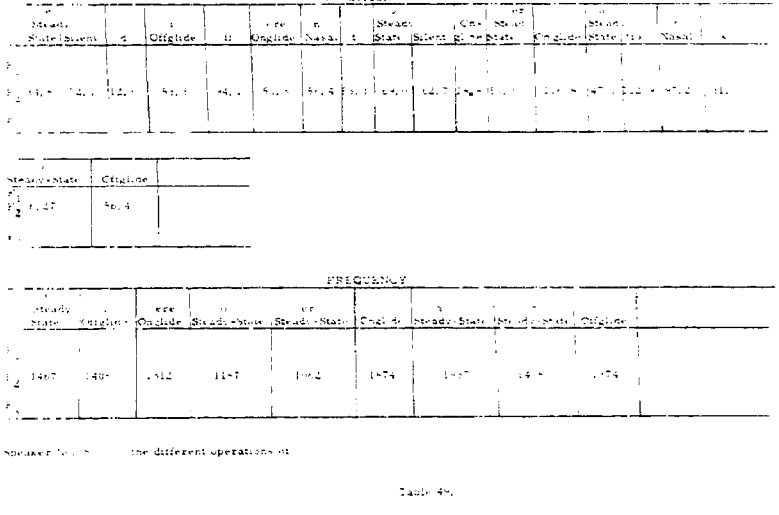


Table 48

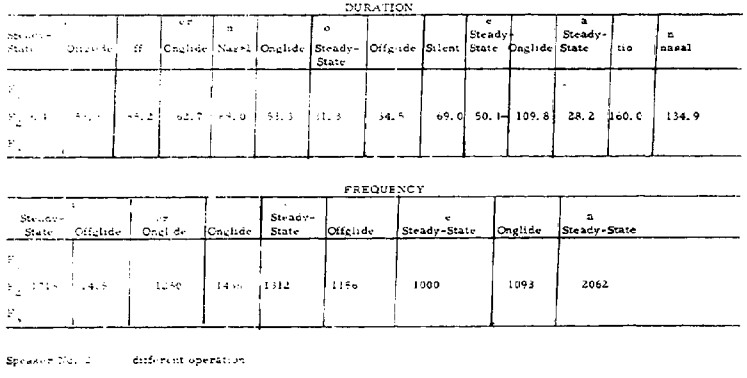
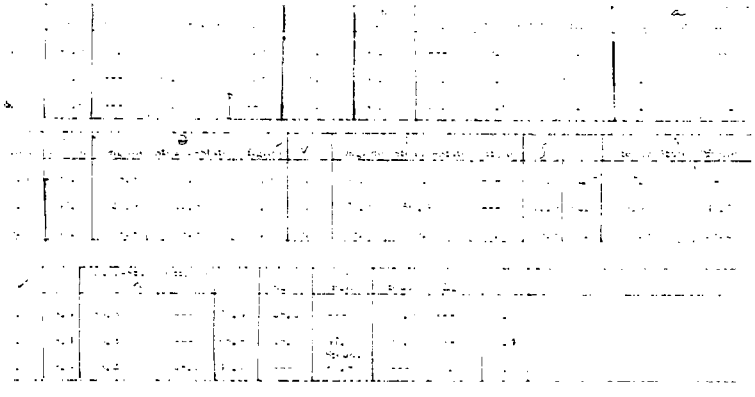


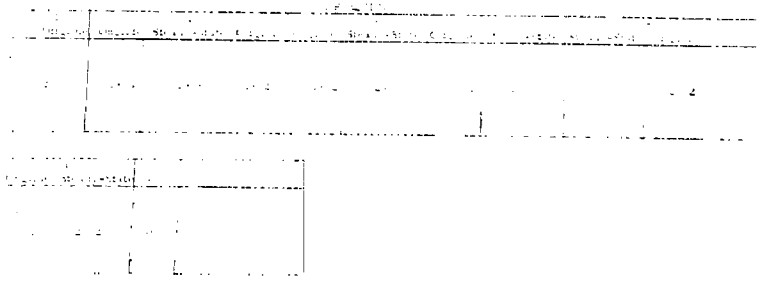
Table 48



							a			a	
					Steady-state	Offglide	Steady-State	Offglide	Onglide	Steady-state	Offglide
						717	717	469	469	531	469
						1250	1250	1156	1562	1656	1562
						2498	2592	2437	2592	2716	2530

										e	I	
					Steady-State	Offglide	Steady-State	Offglide	m	Onglide	Onglide	Steady-State	Offglide
						717	n24	No	No	No	382	No
						1405	1343	1093	1186	1781	2341	No
						2716	2561	No	No	2561	2561	2406

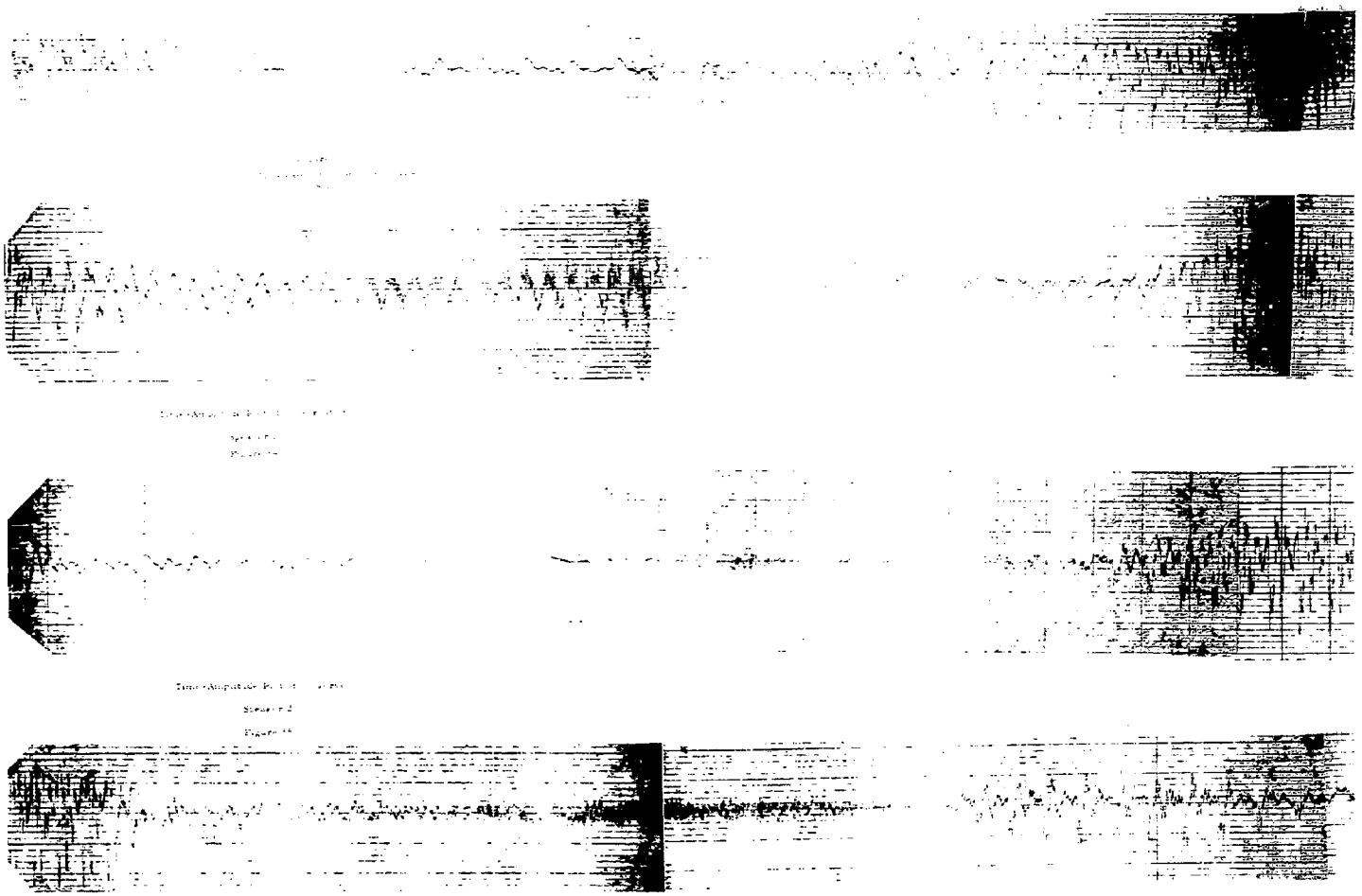
Table 50

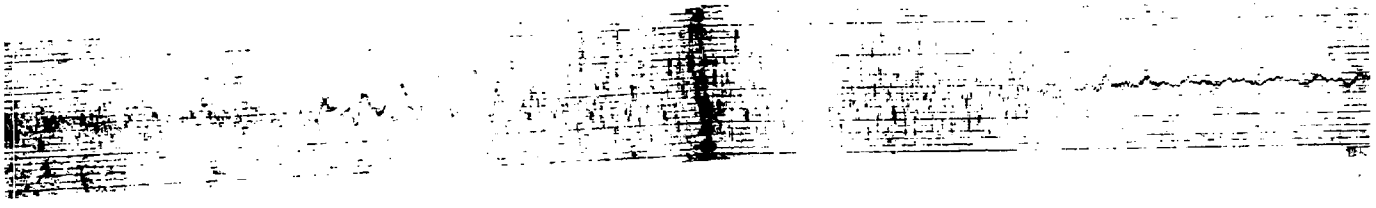


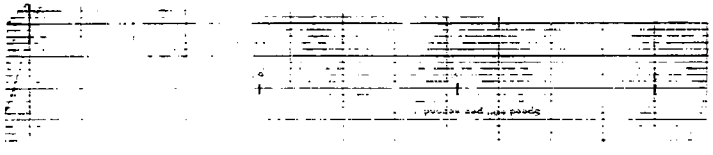
DURATION

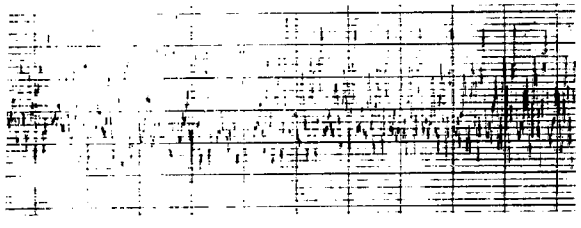
							a			e			
					Steady-State	Offglide	Forward	Silent	Onglide	Steady-State	Offglide	Steady-State	b
							72.1	96.5	28.6	28.0	116.8	28.4	3.3

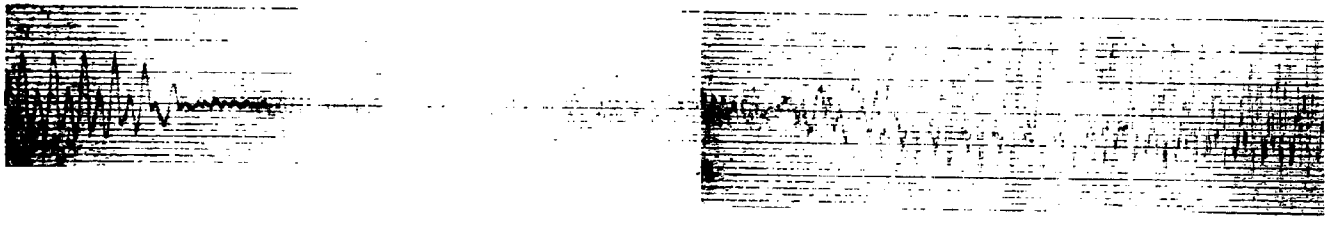
								a				
					Steady-State	Offglide	Onglide	Steady-State				
							75.2		156.9	40.7	72.1	59.6

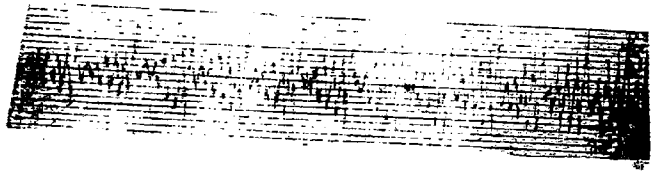












V. FURTHER MEASUREMENTS WHICH INDICATE THE IMPORTANCE OF DURATION AND INTENSITY, AND WHICH SUBSTANTIATE OUR APPROACH.

From time to time we have mentioned that duration, fundamental frequency, and intensity are dimensions of speech which merit detailed analysis of certain minute portions of the speech waveform for effective speech recognition. We have performed such analysis on some of our data (of which Figures 21-46 represent only a portion); from this analysis we obtained a sizeable amount of evidence to substantiate our approach and to indicate the necessity of further study of the dimension of intensity. Some of our results are summarized below.

The importance of intensity measurements is shown by the vowel-like "nonsense" segment of one pitch period duration which follows the m in some in the phrase "of some ancient sages." (see Figure 58). This segment can be ignored by a computer working with the rule mentioned earlier in this section, which specifies the minimal number of pitch periods which are allowed in a legitimate segment.

The necessity of redefining stops is indicated by the spectrograms of different (Figures 47, 48, and 49). In most speakers' pronunciation, there is no stop gap before the t. The definition of stops could thus be modified by specifying that the stop gap may be absent when the t follows a nasal. (This would also apply in a word such as mumps.)

In the words operation and observation the waveform for most speakers shows either no vowel segment or a vowel segment of very short duration between the ʃ and the n in the tion portion. A rule to this effect should be incorporated into the model.

In the observ portion of observe and observation, moreover, (see Figures 36, 39, 42, 50, 51, 52, 53, 54, 55, and 56) the effect of stress or intensity is apparent. Measurements of these two words, spoken by the same speaker, have shown an 11 to 18 ratio in the overall rate of articulation of the same phone classes (observ) in observation and observe; second, there is a 20% variation (i.e. about 200 cps) in the second formant frequency; third, there is a variation in the duration of the individual phone classes (especially er and v) which may be as great as 640% or as little as 7.15%. You will notice furthermore, that our data includes not only spectrograms, but also time-amplitude plots, which have a dynamic range of about 45 dB.

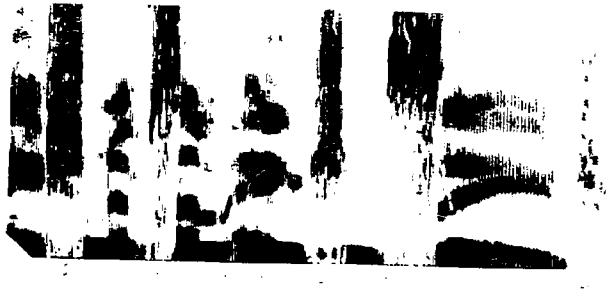
As indicated by the rules of euphonic combination, the t of ancient becomes aspirant in "ancient sage", because the following word begins with s.

In the waveform of "soundness or rottenness" in continuous speech (See Figures 57, 58, and 59) it is difficult to tell whether one r or two were spoken; there is a single r sound indicated, which has an abnormally long duration. Boundaries must be included in the model to specify according to duration whether one r or two are present. Furthermore, a computer program such as that outlined in the following section must be included to provide a contextual means (according to the "correct" or dictionary representation of words) for restoring word boundaries.

The treatment of h as a voiceless vowel or a portion of the vowel segment mentioned earlier in this section is substantiated by the spectrograms of the human mind. (See Figures 62, 63, 64). The i portion of the (Figure 63) has an F_2 frequency of 1529 cycles and the i portion following the h in human has an F_2 frequency of 1405 cycles. There is thus no significant formant change; the h between is merely a voiceless or weakly voiced portion at approximately the same frequency. Furtnermore, in "Mrs. Slipslop," (Figures 65, 66, 67) the variation of the l phone class justifies treating this as a vowel cluster.

Finally, the spectrograms of "which wise sayings" introduce several interesting details. We have two representations of which pronounced by Speaker 1. (Figures 68 and 69). In Figure 68, the duration of the onglide following wh is 69 milliseconds; in Figure 69, the duration of the same portion is 40.7 milliseconds, although the overall duration for the word which is approximately the same in both cases (255.9 ms in Figure 68 and 250.8 ms in Figure 69). Futhermore, in Figure 69, no real steady-state is ever achieved for I, whereas in Figure 68 (with the slower onglide) there is a slight steady-state. The onglide portion thus seems to need corrections and/or normalization: the machine must be instructed, for example, that the more rapid onglide (Figure 69) must be extrapolated, in order to assign the proper frequency, because the steady state frequency in this case (Figure 69) is approximately 100 cycles per second less than in the other instance (Figure 68).

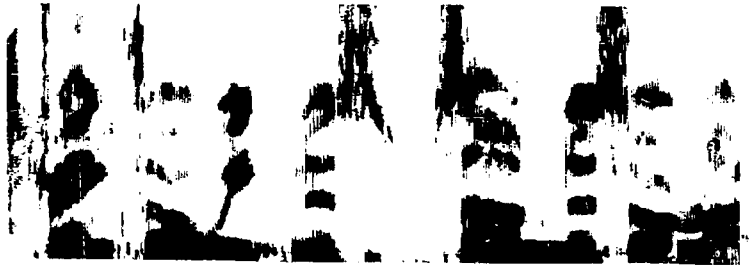
The ch in which also merits attention. Perhaps, as some researchers have suggested, one could sample the ch pattern at some arbitrary point - such as 6 milliseconds after the burst. However, the use of such a technique needs justification, before it can be used without question. Short-time statistics of the ch waveform might be necessary, because of its irregular nature, but certainly overall normalization of which would disproportionately compress the ch portion of Speaker 2's pronunciation, where the duration of ch is only 62.7 milliseconds (Figure 70) as compared to 125.4 and 103.5 milliseconds in Speaker 1's articulations (Figure 68 and 69).



Spectrogram of "... ervation of some ancient sage..."

by Speaker 2

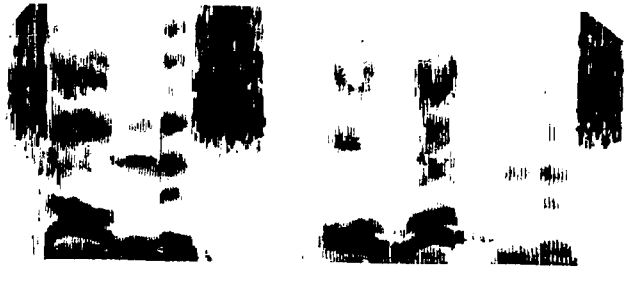
Figure 58

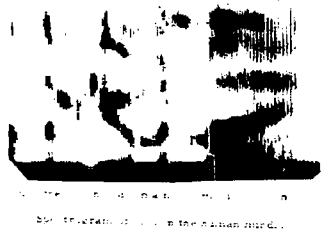


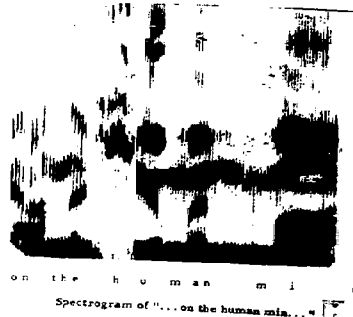
Spectrogram of "... bess soundness or rottenne..."

by Speaker 2

Figure 60



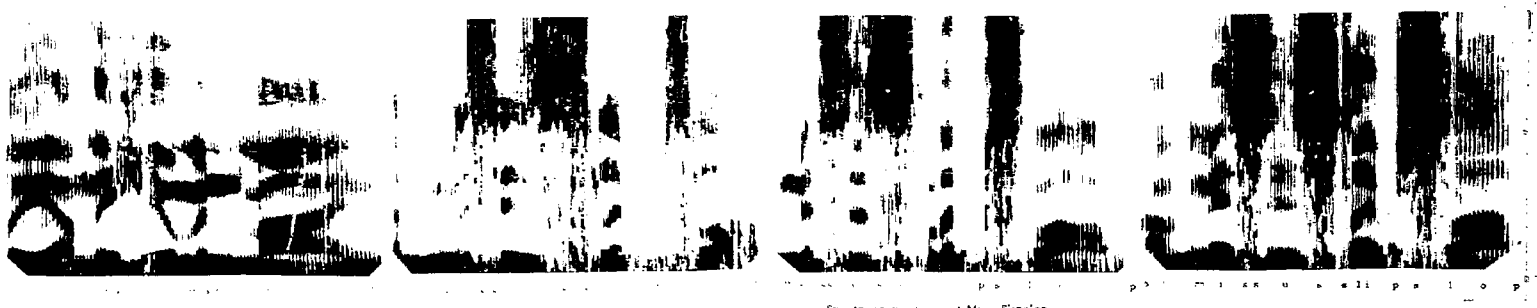




Spectrogram of "... on the human mia..."

by Speaker 2

Figure 63



Spectrogram of "...of Miss Slipslop..."

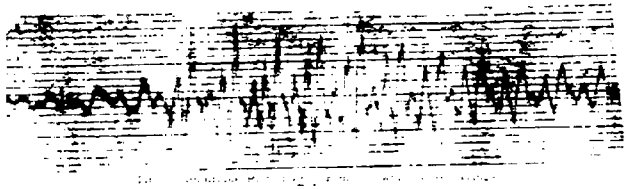
by Speaker 2

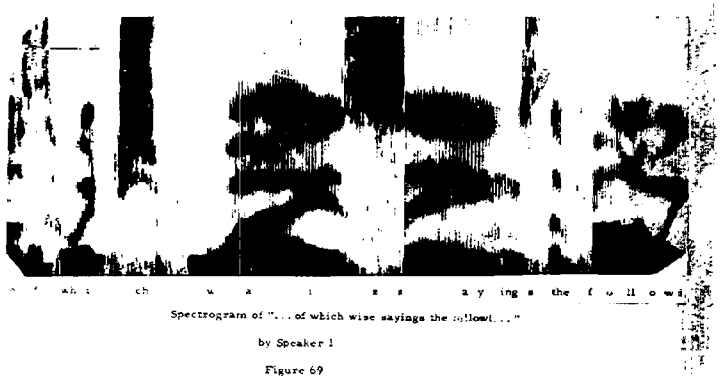
Figure 66

Spectrogram of "...of Miss Slipslop..."

by Speaker 5

Figure 67





Spectrogram of "...of which wise sayings the follow[...]"

by Speaker 1

Figure 69

Spectrogram of "...which wise sayings..."

by Speaker 5

Figure 71

DURATION

	✓	a Onglide	a Steady-State	ʃ	n Nasal	Steady-State	f	e	Fi. St.	Nasal m	Discontinuity
F_1											
F_2		34.5	47.0	122.[illegible]	[illegible]	[illegible]	[illegible]	73.2	60.7	56.4	[illegible]
F_3											

	Onglide	Steady-State	u	ʃ	Silent	a	Onglide	a Steady-State	Silent	z
F_1										
F_2	[illegible]	[illegible]	50.1	94.1 21.4	47.4	1[illegible]7.72	147.4	147.4	60.7	72.[illegible]
F_3										

FREQUENCY

	a Onglide	a Steady-State	o Steady-State	o Steady-State	a Onglide	a Steady-State	a Onglide	a Steady-State
F_1								
F_2	1437	2011	1494	1467	1781	2093	1467	2166
F_3								

Speaker No. 1. [illegible]vation of some ancient sage. Table 57

DURATION

	s	Silent	t	r Steady State	Onglide	n Nasal	κ	m	Onglide	Steady State	r	a	Onglide	Onglide	Steady State	Silent	n Nasal
F_1																	
F_2	47.5	53.3	25.3	34.5	57.[illegible]	62.7	47.[illegible]	[illegible]	62.7	[illegible]	[illegible]	41.7	[illegible]	[illegible]	21.[illegible]	[illegible]	34.5
F_3																	

	Onglide	e Steady-State	Offglide	ss	Silent	z	Onglide	Steady-State	Silent	n Nasal	Silent	Steady-State	ss	Onglide	r Steady-State
F_1															
F_2	15.[illegible]	25.2	41.[illegible]	125.4	144.9	[illegible]	125.6	[illegible]	140.7	25.2	[illegible]	[illegible]	46.2	[illegible]	41.[illegible]
F_3															

	Onglide	m	n Nasal	Steady-State	ss
F_1					
F_2	[illegible] 57.2	44.1	133.[illegible]	62.7	72.1
F_3					

Speaker No. [illegible] Strength or weakness soundness or rottenness

Table [illegible]

DURATION

	ɛ Offglide	ʌ	e	S	n Nasal	O Steady-State	O Offglide	f	s	o Onglide	o Steady-State	o Offglide	m Nasal
F_1										34.5	21.9	43.9	
F_2	65.[illegible]	72.2	65.2	155.8	72.1	31.3	56.4	56.4	128.6	No	40.7	62.7	103.5
F_3													

	Silent	æ Onglide	æ Steady-State	n Nasal	ʃ	ʌ	n Nasal	Silent	s	a Onglide	a Steady-State
F_1											
F_2	76.4	63.7	53.3	50.1	109.[illegible]	50.1	94.1	53.3	188.2	134.9	87.8
F_3											

FREQUENCY

	Offglide	e	o Onglide	o Steady-State	o Offglide	o Onglide	o Steady-State	o Offglide	a Onglide	a Steady-State	ʌ
F_1			531	624	No	428	717	655			
F_2	1281	Diphthong	No	1436	1216	No	1218	1093	1968	2417	1467
F_3											

	Onglide	a Steady-State
F_1		
F_2	1498	2031
F_3		

Speaker No. 2. [illegible]servation of some ancient sage. Table 58

FREQUENCY

	r Steady State	Onglide	o Onglide	o Steady-State	Onglide	ea Onglide	ea Steady State	e Onglide	e Steady State	e Offglide	o Onglide	o Steady State	e Steady State
F_1													
F_2	1374	1324	1312	1031	1031	2000	2248	1593	1656	1593	1436	1281	1467
F_3													

	r Onglide	r Steady-State	o Onglide	e Steady-State
F_1				
F_2	[illegible]	1124	1167	1529
F_3				

Speaker No. 1 Strength or weakness soundness or rottenness

Table 59

FREQUENCY

[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	Steady State	o Onglide	Steady State	r Offglide	o Onglide	Steady State	Onglide
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1624	1374	1000	1124	1000	1343	1436

Steady State	Onglide
[illegible]	[illegible]

Table 60

DURATION

	ð	r	ə		h			u	m	a	n	m	a On-glide	a Steady State	I Off-glide	n	d
F_1					72.[illegible]			87.0									
F_2	[illegible]	75.4	21.9	47.0	47.0	21.9	47.0	47.0	56.4	28.2	56.4	37.6	28.0	59.6	116.8	37.6	Not Clear
F_3																	

FREQUENCY

	ð	ə		ə	Offglide	Onglide	a Steady-State	I Offglide
F_1								
F_2	[illegible]	[illegible]	[illegible]	1937	1374	1124	1374	1761
F_3								

Speaker No. 1 [illegible] the human hand.

Table 62.

DURATION

	Onglide	Steady State	Offglide	Onglide	Nasal	Offglide	Onglide	Steady State	Offglide	Nasal
F_1										
F_2										
F_3										

FREQUENCY

	Onglide	Steady State	Offglide	Onglide	Nasal	Offglide	Onglide	Steady-State	Offglide
F_1									
F_2									
F_3									

Speaker No. 2 on the human mind

Table 63

DURATION

	Onglide	Steady State	Offglide	Onglide	Nasal	Offglide	Onglide	Nasal	Onglide	Steady State	Offglide	Nasal
F_1												
F_2				90.1	78.4		40.7	78.4	53.3	85.2	97.2	85.2
F_3												

FREQUENCY

	Onglide	Steady State	Offglide	Onglide	Nasal	Offglide	Onglide	Steady-State	Offglide
F_1									
F_2	1150	2000	1187	1436	2062	1000	1529	1150	1281	1906
F_3									

Speaker No. 5 IV on the human mind

Table 64

DURATION

	Nasal	Steady-State	Steady State	Onglide	Steady State	Silent	Silent	Onglide	Steady State	Offglide
F_1										
F_2	93.1	21.9	97.2	40.7	100.4	25.1	21.9	15.6	85.2	94.1	34.5	69.0	59.6	40.7
F_3														

	Silent	Burst
F_1		
F_2	94.1	3.13
F_3		

FREQUENCY

	Steady-State	Steady-State		Onglide	Steady-State	Onglide	Steady-State	Offglide
F_1								
F_2	1624	467	000	1250	1436	965	1124	1062
F_3								

Speaker No. 2 Mrs. Shipslop

Table 66

DURATION

	nasal	Steady State	Onglide	S. S.	Silent	Onglide	S. S.	Silent
F_1								
F_2	142.4	88.2	93.1	37.6	88.2	37.6	94.1	3.13	69.0	3.13
F_3								

FREQUENCY

	Steady-State	Onglide	Steady-State	Onglide	Steady-State	Offglide
F_1						
F_2	1467	1429	1718	940	1462	1000	1762
F_3						

Speaker No. 1 Mrs. Shipslop

* Discriminate.

Table 65

	Steady State			Steady State	Silent									
	40.7	103.5	43.9	15.9	66.7									

	Steady State			Silent		Silent	Steady-State					
	103.5	60.4	80.1	34.5	61.3	78.4	65.5					

FREQUENCY

	Steady State		Steady State		Steady State		Steady State					
	1124	872	1760	1093	1282	1447						

Speaker No. 1 of which wise savings the following

DURATION

		Steady State	Silent		Silent		Onglide	Steady State						
		65.0	28.2	61.3	125.4	40.7	34.4							

FREQUENCY

		Steady State									
		685	1474								

Speaker No. 2 which wise savings

DURATION

FREQUENCY

DURATION

FREQUENCY

Which conforms also to our segmentation principle as outlined in Section 5 of this report. Whi is one consonant-vowel portion; chwa(i) is another.

These are not exhaustive examples, but only a few significant details which justify our classification, and as we shall see in the following section, which also support our approach to segmentation.

SECTION 5: SEGMENTATION AND CONSIDERATIONS FOR COMPUTER OPERATIONS

INTRODUCTION

Our examination of the linguistic, phonetic, genetive, and acoustic aspects of speech has substantiated our original concept of the multi-dimensional model as an orderly basis for representing speech information, and has somewhat modified our original representation.

Furthermore, our research has enabled us to develop a method of segmentation which is suitable for automatic speech recognition. This segmentation principle is explained in Part I of this section. Moreover, as we mention in that discussion, we have tentatively applied this technique to spectrograms of words from Visible Speech , Truby, and also to spectrograms and time-amplitude plots of discrete and continuous speech which were generated during this project, in order to ascertain whether our approach to segmentation seems warranted by the evidence. Such evidence seemed necessary to substantiate the assumptions (linguistic and genetive) which were made not only for the model, but also for the rules of euphonic combination and coarticulation. Our measurements have verified both our approach and our method of segmentation on an acoustic level. Thus we believe we have selected those segments of speech which best describe the realities of speech events, yet which will be most meaningful to an automatic speech recognizer. In so doing, we have successfully integrated the data available from the various sources -- genetive, linguistic, phonetic, acoustic, etc. -- into an orderly representation which could serve as the basis for a general purpose recognizer.

In Part II of this section, we outline several possible approaches to the computer program which is to resolve the perceived sounds -- i. e. to perform the dictionary match. Part II completes the study by providing an outline of the functions of the various phases of the recognizer and a discussion of how our contributions may be used in each phase to make the recognition program operative.

I. SEGMENTATION

Throughout this project, we have emphasized that the place and manner of various consonants can change, according to the vowel which precedes or follows it; thus the genetive, linguistic, and acoustic representation of a consonant may change. As research progressed, moreover, it became apparent that certain sounds which had traditionally been described as separate entities were, as H. M. Truby emphsizes, acoustically interdependent, or coarticulated. This evidence of coarticulation, and also

the evidence of transitions provided by Haskins Laboratories in their attempt to produce synthetic speech, have greatly influenced our concept of a meaningful machine segment. On the basis of the evidence we have examined, we recommend that, generally speaking, the most meaningful unit for machine recognition will be a "consonant-transition-vowel" segment, including any offglide of the sound which precedes the consonant and thus helps to identify it, and including the onglide of the vowel to the point where a steady-state is achieved.

Our method of segmentation can be illustrated by comparing it with the segments proposed by Gunnar Fant and Bjorn Lindblom in their "Studies of Minimal Speech Units." In Figures I-1 of that article, the authors have marked 18 segments in a spectrographic record of the words "Santa Claus." Segments 9-15 of their analysis would be treated as one segment in our model. This segment would include the k (which first shows up in the offglide pattern of the vowel _) , the l (which is coarticulated with the k), and the onglide and steady-state portions of the ɔ : sound. From the middle of the steady-state to the end of the z forms another segment -- a vowel-consonant combination. Such a segment provides a meaningful unit for machine recognition, since it depends upon the rate of transition from the consonant to the vowel (and vice versa) rather than the absolute formant frequency values, which may change according to their environment and context. Segments are "matched" by correlating the smallest articulated acoustic representations of these segments. In order to match them more perfectly we can either (1) equalize them, by changing the duration of portions of a spectrogram, without altering their spectral density characteristics, or (2) we can derive short-time statistics to compare two portions of time-amplitude plots.

Our method assumes that the vowel and consonant components of speech are interdependent and should not be separated in recognition. To justify this assumption, we have performed considerable measurements of words in Visible Speech . If our technique applies to these to these words, it should apply to most samples of English speech. We are emphasizing the interdependence of sounds in continuous speech, and the words in Visible Speech are discretely articulated speech. We found in Visible Speech considerable evidence to substantiate our principle of segmentation. Furthermore, we performed measurements on our own samples of continuous speech, and found that the segmentation principle was equally applicable.

Although our study is not directly concerned with vowels, we have found that the rate of change found in the vowel onglide provides a valuable indication of what consonant preceded that vowel. For instance, the onglide

of i in fee has a duration of 82 milliseconds, whereas the duration of the onglide of the i in key is 38 milliseconds (Visible Speech, pages 121 and 51). Furthermore, in week, the duration of the i onglide is 57 milliseconds, whereas in he it is 304 milliseconds (Visible Speech, pages 207 and 113). To account for this, we can set limits for the rate of transition, so that beyond those limits, the sound must belong in another category. That is, if the rate of change were below a certain slope, the sound would be matched with one segment; but if the rate of change were above (i. e. more rapid) than that slope, the sound would be classified in another category.

This rate of change is responsible for the duration of the vowel onglide, but also of the steady-state of the vowel F_2 frequency. This, too, seems to vary according to the consonant which precedes the vowel; in leave, the steady-state frequency of the i is 2099 cycles per second; in reed it is 1808 cycles per second. This change in frequency according to the consonant will obviously influence the slope of the consonant-vowel transition. For this reason, it does not seem feasible to normalize the vowel steady-state, and still expect accurate recognition. Instead, we have specified a segment which can readily accommodate the wide variations in duration and frequency characteristics which our measurements have found.

It may seem that our coarticulated sound cluster is the rough equivalent of what is commonly called a syllable, such an analogy is not inherent in our thinking. Instead, we have developed our principle of segmentation from linguistic, genetive, phonetic, and acoustic aspects, and have sought continuously to specify the smallest recognizable (and therefore constant) articulated unit of sound. Our basic machine acoustic unit, should not be considered the equivalent of a syllable.

A main source of our segmentation principle was the information derived from our linguistic study -- particularly the grammar of Sanskrit. For in that grammar, an individual phone class is specified to represent each consonant-vowel combination; these classes of sounds have been tested to determine their applicability to the English language. It was found that acoustically, English speech can also be divided into classes of "C-V" combinations, although the classes are not the same in each language.

A. Consonant Clusters

The English language is not merely a sequence of CV and VC combinations: two other important groups of sounds occur -- vowel

and consonant clusters. While this study does not attempt to deal with the special problems presented by vowels, we do have certain recommendations about the treatment of consonant clusters.

For the purposes of the perceiver (for which we have developed this segmentation principle) most of what are commonly regarded as clusters will be treated as separate entities. A consonant cluster such as str seems to be different, acoustically, from s + t + r. By way of evidence to justify this position, it has been found that the t in treat may possibly be aspirated whereas it is highly unlikely that the t in street will be aspirated. Similarly, the r in trade may have a shorter duration and a higher F_2 frequency than the r in raid. Thus str or st or tr is not the mere sum of its components, but a special class of sounds which requires certain movements of the articulators, and which thus produces a distinctive acoustic pattern.

B. Refinement of the Concept of Coarticulation.

Our measurements of the material in Visible Speech yielded evidence to substantiate the concept of coarticulation. There was no measureable voiced onglide between the p and e in person, (p. 180), between p and e in pep up, (p. 85) and between p and i in pipe (p. 85) Also, in pep (p. 84), pass (p. 139), and pup (p. 101), there is a close correspondence between the frequency at the start of the voiced onglide and the (F_2) frequency of the vowel steady-state.

The evidence presented by H. M. Truby yields even more examples of coarticulation than he points out. (Acta Radiologica, Supplementum 182, Stockholm, 1959.)

(1) For instance, the spectrogram of the word jaunt (p. 19) shown no stop between the n and the t -- as we had also noticed in the word different in our own data.

(2) Furthermore, as we mentioned in Section 4 of this report, the r in words such as cheer (p. 14) and George (p. 19) will be treated not as a separate class of sound, but as a vowel cluster or portion influencing the offglide characteristics of the accompanying vowel.

(3) In jounce (p. 20) Truby has used two phonetic symbols to represent the vowels a and u. We would instead make this combination a special class of vowel.

(4) He has also represented the ce portion of jounce phonetically as ts. The sound here, we think, is more than a sequence of t and s: again, it seems instead to be a unique class of sounds.

(5) In a pl combination, such as in plink (p. 20), we might have to specify that the p can be unaspirated. In blink (p. 20) it is possible that the b is modified by the l, so it might be advisable to include bli as a distinct category; furthermore, the "typical" energy distribution of k may be altered in kl combinations, such as clip (p. 25). Perhaps even gl (as in glib p. 28) must be treated as a separate acoustic element.

(6) Again, y, r, l, and w seem often to be coarticulated with the adjoining vowel; thus words such as tweak (p. 47) are only one CVC utterance -- the w becoming part of the following vowel.

(7) Finally in the kl combination in the word sclaff (p. 51), there is a strong possibility that the k will not be aspirated, when the following vowel is emphasized. This is much like the case of pl in plink mentioned above (Truby, p. 20). This might also be true of the p in spree (p. 52) and in other cases where the vowel following the p is emphasized.

These examples of coarticulation occur within individual words, it is also highly possible that coarticulation may occur at word boundaries, as in the tl of "it lasted" mentioned in Section 3 of this report.

Coarticulation seems an important concept in describing the realities of speech events. The illustrations used above are not a formal organization of all possible incidences of coarticulation, but they point out certain "problem" segments or combinations which must receive special attention. Our principle of segmentation is designed to deal with just such problem segments as these, by using CV units, and also by treating such sounds as y, r, l, and w as vowels or members of "vowel clusters."

II INFORMATION ON THE OCCURENCE OF RULES

Using our data on [illegible] speech we studied the possibility that the rules of euphonic combination were indeed operative in certain segments of speech samples of some of the subjects but that these very same segments could indicate that these same rules were not operative in the speech samples of the rest of the subjects whose speech was analyzed for this study. We include it in Table 6 because although inadequate in any conclusive sense, it provides some initial information about occurrence frequency that might influence our ordering of our euphonic combination rules in the computer. We eventually plan to order our rules so that those rules that apply most often come first. We are grouping our rules by related situations of sequence and ordering them by groups. Statistical information on the probability of operation of one or more of these rules could possibly improve the efficiency of our program.

Table [illegible] Information on the Occurrence of Rule

	The Word	Segment of Speech Regarded	Rule	Speaker I	II	III	IV	V
1.	It	"it is"	t becomes d	Yes+	Yes	Yes	Yes	Yes
2.	Observation	'observation'	b voiceless	No	No*	No	Yes	No
3.	Sage	'ancient sage	s voiceless		Yes	Yes	Partly	Partly
			glottal stop before ancient		Yes	?	Not Available	
			no stop gap for t	Yes	No	No	Not Available	
4.	Human	'on the human mind'	no transition for 'n' (i.e. dental n)	Yes	?	Not Available		Yes
5.	Weakness	'weakness soundness'	2 s's become 1 s	No (Pause)	No	No (Pause)	No (Long s)	No (Pause)
6.	Soundness		no d	Yes	Yes	Yes	?	No
7.	Himself	'give himself'	no h	No	?	?	?	No
8.	Some	some pains	m not released	Yes	Yes	Yes	?	No
9.	Pains		p voiced (aspirated)			Yes	No	No
10.	Different	'different operations'	[t] coarticulated with [o].	Yes	Yes	?	No	?
			t voiced	Very Short	Yes	?	No	?

+The rule does apply

*The rule does not apply.

Table 72. Information on the Occurrence of Rules

	The Word	Segment of Speech Regarded	Rule	Speaker I	II	III	IV	V
11.	Mrs. Slipslop	"Mrs. Slipslop"	2 s's become 1 s	No (Long s)	No (Long s)	No (Long s)	No (Half-longer)	
12.	Wise Sayings	"Wise Sayings"	s becomes z	No	No	?	No	No
			z becomes s	Partly	Mostly	?	Mostly	Partly
			2 s's become 1 s	No (Long s)	No (Long s)	?	No (Long s)	No (Long s)

Speaker II long s of "wise sayings" less than half as long as long s of "Mrs. Slipslop", somewhat longer than single s of "Slipslop."

Speaker III s of "wise sayings" very little energy.

III OUTLINE OF APRROACHES TO THE COMPUTER PROGRAMS

The sequence of speech sounds in the construction and transmission of words and utterances is due to the physical limitations of speech-producing mechanisms and to the demands of linguistic tradition. We have developed rules of euphonic combination, based on an understanding of preferred positions for sound, to determine how sounds are modified by speech-environment. In the conceptualization of a multi-dimensional model for speech recognition, we integrated data on the genetive, phonetic, phonemic, and acoustical aspects of speech -- in a manner faithful to the realities of speech events. The rules we derived from this and other information represent the first time an orderly approach to the modifications of adjoining phone classes has been clearly defined. And the rules are practicable. We have reduced them to symbolic representation and prepared them for use in a computer program. We have approximately five hundred rules; but we were able to group these to reduce the number of rules the computer must store. This grouping was made possible by the nature of the structural ordering of phone classes. Phone classes are related by the dynamics of articulation: p, t, and k are related, as are g, d, and b. So, that which applies - descriptively to the combination of k and g applies also to z and d or p and b. Therefore, a computer need only store about fifty rules for defining the effects of adjoining forms on each other.

We can choose from a number of methods in designing the system by which our machine actually computes what euphonic reductions it must account for. At present, three such methods are under consideration. In each, the application of our rules is fundamental.

The reversed rule method involves the application of all applicable reversed rules to any given situation. Predetermining possible consonant reductions will, to an extent, mitigate the formidable problem of such an approach (the proliferation of possible rule applications.) Constant reference to a list of allowable consonant clusters after each rule application is still admittedly inefficient. So of our three methods, the reversed rule method is the one we are least likely to employ.

The consonant cluster method involves the construction of a dictionary containing all reduced forms of consonant clusters and all possible antecedents, of those clusters. By first finding all the correct antecedents of each initial cluster, we establish the environment for any terminal cluster we consider. Of course an understanding of antecedents (unreduced consonant clusters) requires an understanding of how consonant clusters are reduced in speech. So it is impossible to construct lists without our rules of euphonic combination. Once such lists are established

for the determination of terminal-initial clusters we may apply them to medial clusters: breaking medial clusters into terminal-initial clusters and then solving. But at present the problem of medial cluster segmentation, among other problems, makes it more likely that we will use an alternative solution (by our rules).

In our treatment of consonant clusters, we considered the treatment of semivowels. These we found it easiest to deal with by the rules alone - that is, without the implementation of lists describing particular or even general occurrences of the semivowel in speech. The rules are in this case sufficient to account for the elision or insertion of a semivowel.

The Reduced Word Dictionary Method is similar to the consonant cluster method in that both depend on a thorough and comprehensive application of the rules of euphonic combination to provide a listing of reduced forms. Since here we are dealing with whole words, word division is a primary concern: our rules are of further use since they represent initial breakthroughs in the treatment of the problem of word juncture. Furthermore we are now evaluating a number of techniques to facilitate the placing of a word division. Particular attention is given to technique such as alphabetizing search arguments, either in context or isolated from context. Protection and verification techniques are also in the process of final formulation (these latter may be used with either the consonant cluster or the reduced word approach.)

We have not yet come to a final decision about the particular method we will choose. To do this would require a decision about the kind of computer machinery we will employ. Relative differences in the amount of clerical work necessary in the compilation of different dictionaries will also require scrutiny. And after that, we will have to make tests on computers to compare time differences in the methods with all their accompanying techniques.

In our original proposal we wrote: "With the recognizer, however, the problem is not to discover words (these being known in advance to the designer), but rather to ensure that borders are included properly in the machine output as spaces between words. That is, a machine that operates with acoustical data must make decisions about non-acoustical phenomena. This problem is no doubt beyond the capability of present theory, nor does its solution seem especially urgent in the context of other, more basic considerations. However, its relation to some other problems may bring it in for cursory study during the proposed research program. The work above shows that we have gone far beyond this.

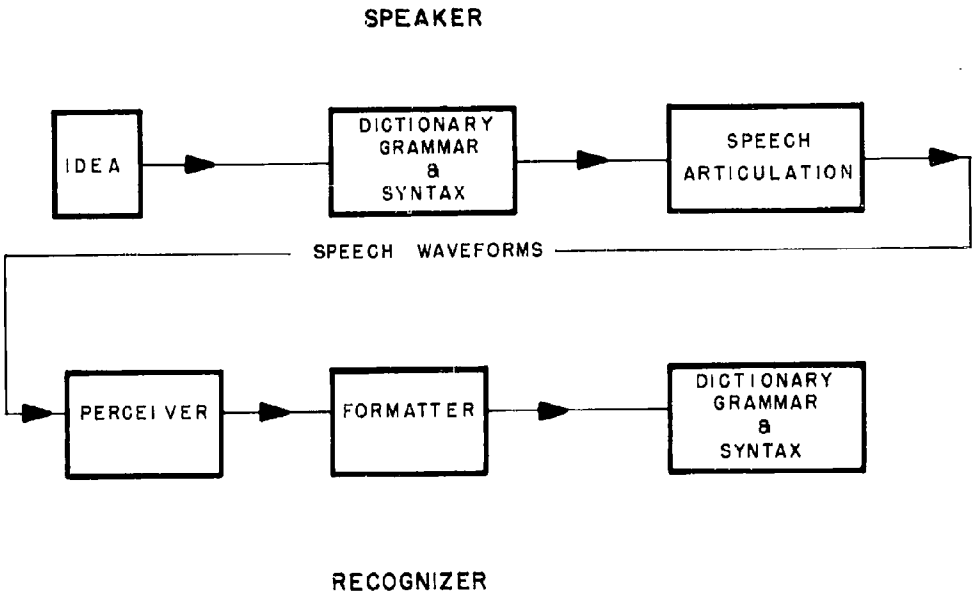


Diagram of Speech Production
and Perception Process

Figure 72

SECTION 6: CONCLUSIONS

Figure 71 is a schematic diagram of the speech and recognition process. Our study has centered about the speech articulation phase, which obviously bears a direct relationship to the nature of the speech waveforms. Since it is the speech waveforms which comprise the input data to the perceiver and hence the formatter, we must understand the possible and probable speech events which occur in the articulation phase before an automatic recognizer can be designed.

We are convinced that the acoustic information which can be gathered from continuous articulation is more complex than a mere succession of phonemes. We have understood this complexity to consist of slurs, or the incorrect pronunciation of certain phone classes which occur in the orthographic form of language.

To explain this imprecise pronunciation, we have collected a large body of data which we have organized into more than 500 rules of euphonic combination. We have found that group theory can be employed to order these rules according to the degrees of freedom available in the articulation of speech sounds and to compress this body of rules into 50 rules in symbolic notation, suitable for computer storage. Such stored information provides an error-correcting code which can be used to reconcile imperfectly articulated continuous (and again we emphasize, normal) speech with orthographic script.

Speech recognizers have been built in the past which assume that a machine is capable of recognizing the words or phonemes. Most of these machines (See Figure 18, Section 4) have enjoyed only limited success. In all these designs, the vocabulary has been limited; moreover, when the phoneme was the segment to be recognized, the single phoneme had to be articulated in a fixed-consonant environment.

It is desirable to extend the success of these methods beyond their present limitations: that is, it is desirable to extend the size of the vocabulary recognized and the environments in which sounds can be recognized. It has often been possible, though, that by refining the existing methods one could increase the number of words recognized and so extend the applicability of the present speech recognizers.

Perhaps, as we have suggested throughout this report, the more useful approach is not to be found in attempting to develop a speech recognizer with the limited amount of information which is presently available. The more useful approach might be to design a method of computer operation which can anticipate and account for acoustic imprecisions of speech. With this as a goal, we have examined the nature of continuous ("normal") speech,

in an attempt to ascertain what imprecisions of articulation can be expected. Our study, we believe, has been conclusive, if not exhaustive, the significant details of the speech waveform which were examined in Section 4 clearly demonstrate the validity of our approach. Certainly extensive proof would require the generation of additional data.

The most outstanding characteristic of continuous speech, and that which most clearly distinguishes it from discrete speech, is that in continuous speech sounds modify surrounding sounds, in a continuous series of events which are neither a multiplication nor an acceleration of the events of discrete articulation. We have compiled extensive evidence which clearly demonstrates that: (1) the articulation of words or vowel sounds in isolation results in waveforms which are significantly different from the waveforms of the same phone classes spoken in continuous speech. (2) two or more phone classes tend to be coarticulated (spoken as one sound). The coarticulated sound has a waveform which is significantly different from the waveform of either and/or both the component phone classes in careful articulation. (3) the word boundaries which are found in orthographic script are almost totally lost in continuous speech.

It has been our contention that these combinations or modifications of sounds occur in the English language in a predictable way, which can be accounted for according to determinate rules; furthermore, these rules can be related to one another in an orderly fashion. On this basis, a model can be constructed which is patterned according to the various dimensions of sounds -- place and manner of articulation, degree of resonance and aspiration, intensity, duration. Such a model thus would be called "multidimensional."

Conceivably, this study could have been undertaken by attempting to collect vast samples of presentday spoken English. We have chosen instead to begin where more evidence is more readily available. We have at our disposal, for instance, a dictionary of the English language, which lists in phonetic symbols the accepted pronunciations of each word. A large body of knowledge, the result of thousands of years of linguistic study, is equally available: this linguistic literature thoroughly describes the sound changes which have occurred in the historical development of languages (for example, the German d became the English t). By applying the Ergodic Theory from physics, we were tentatively able to assume that these historic examples of sound change might provide a basis for the kinds of sound change which occur in spoken language today, since all the Proto-Indo-European languages studied utilize the same physical modes of production. We then proceeded to test the body of Rules for Euphonic Combination on presentday speech. Certain of these rules found justification; others were modified or rejected, according to the evidence.

Moreover, the sandhi rules of Sanskrit describe modifications or substitutions of certain phone classes, when the phonetic environment is altered. This is much the same phenomenon as what we have pointed out in our rules of coarticulation and euphonic combination. In the present study, it was thought that the speech waveform of certain words in English might contain phone classes which exist in the spoken language, but which are incorrect according to the orthographic indications. Certain examples of this phenomenon have been cited in this report: for example, bet you often becomes be chyou in continuous speech.

In the Multidimensional Model the classification of a particular sound depends upon the degree of freedom which is available in the physical process of sound production. Sounds which are "adjacent" in the model are sounds which are produced almost identically. As indicated on the diagram of the model included as part of Appendix B, the horizontal axis represents the degree of aspiration or resonance, the vertical axis represents the place of articulation, and the depth axis represents the manner of articulation. With such a method for ordering speech sounds, we can conceive of computer programming which depends conceptually on the perceiver to replace unacceptable phone classes with the phone class whose waveform characteristics are nearest to the "incorrect" class which was presented. Thus in the case of bet you becoming be chyou, it is recorded as a rule of euphonic combination that the alveolar stop t becomes the palatal affricate ch before the semi-vowel y. This rule can be used in connection with the model representation to anticipate and correct imprecise articulation, as is found in continuous speech.

Furthermore, the segments which we have described are much more flexible than any previously mentioned by other researchers. The use of a C-V combination as a basic acoustic segment allows both for individual variations in the pronunciation of certain phone classes, and also the acoustic variations which result from the phonetic environment of a given phone class.

Clearly such a system as we outline above places less restriction on the person who uses this machine: he is no longer limited to the number of phonemes or the environment which can be allowed; moreover, the articulation need not be strained -- normal speech can be correctly perceived and printed.

APPENDIX A

These data are a very small portion of a corpus or words and sentences transcribed from the speech of a native speaker of Vietnamese. We believe that the transcription is accurate and the data are sufficiently complete for this analysis. We do not, however, know what dialect the informant spoke, and it is possible that this analysis is not valid for other dialects.

The following list of words shows all the stop consonants which occur in final position in this dialect. All final stops are unreleased; that is to say, there is no stop burst or aspiration. The symbol $\begin{bmatrix} k \\ p \end{bmatrix}$ represents a consonant articulated with two complete simultaneous closures. One is at the back of the mouth where [k] is articulated and the other is at the lips where [p] is articulated. The symbol [ɑ] represents a low back unrounded vowel; the symbol [ɔ] represents a somewhat diphthongized mid-back rounded vowel. The symbol [:] means that the preceding vowel is long.

p

t

k

$\begin{matrix} k \\ p \end{matrix}$

The above list shows that we have four phonetically different

final stops; the problem is to decide how many different phonemes there are. This means we must establish which phonetic differences are relevant (i.e. word-differentiating) in this language. The phonetic differences are:

(1) The difference between [p] and [t]

(2) The difference between [p] and [k]

(3) The difference between [p] and [$\overset{k}{p}$]

(4) The difference between [t] and [k]

(5) The difference between [t] and [$\overset{k}{p}$]

(6) The difference between [k] and [$\overset{k}{p}$]

The best method of establishing the fact that the differences between two phone classes are relevant is to find a minimal pair. A minimal pair consists of two words which are identical in every speech sound except one. If a native speaker says that the two members of this pair sound different then the phones which are different in the two words belong to different phonemes.

In the above list, we have only one minimal pair - [tʌop] and [tʌo$\overset{k}{p}$]. The existence of this pair tells us that [p] and [$\overset{k}{p}$] do not belong to the same phoneme.

Although there are no other minimal pairs, there are some near minimal pairs. A near-minimal pair is a pair which is identical in some segments and different in others. In using a near-minimal pair to do this phonemic analysis, we make the assumption that the differences between the final stops of two Vietnamese words are independent of the differences between the initial consonants of these words. There is always some risk involved in making such an assumption, but if we do not make it, we cannot continue the analysis.

The assumption is bolstered by the fact that no language has yet been analyzed in which differences between final consonants depend on differences between initial consonants. We will assume, then, that the differences between the [t] of [ʃ ∝ t] and the [k] of [t_y ∝k] are not dependent on the differences between the initial consonants [ʃ] and [ty]. We decide that [t] and [k] belong to different phonemes because they appear in the same position, in final position after the vowel [∝].

In order to compare the final stop of [f∝:p] with those of [ʃ ∝ t] and [t_y ∝k], we must make the further assumption that the differences between the final stop of [f∝:p] and those of [ʃ ∝ t] and [t_y ∝k] are not dependent on the length of preceding vowel. Again, we would rather not make assumptions like this, but there is no help for it. Having made this assumption, we compare the final stop of [f∝:p] with that of [ʃ ∝ t] and conclude that they belong to different phonemes. Likewise we compare the final stop of [f∝:p] with that of [t_y ∝k] and conclude that they belong to different phonemes.

Out of the six possible comparisons which we listed earlier, we have carried out four. We have established that the following pairs of stops cannot belong to the same phoneme.

(1) [p] and [$\overset{k}{p}$]

(2) [p] and [t]

(3) [p] and [k]

(4) [t] and [k]

The fact that [p] contrasts with [t] and [k]; and [t] and [k] contrast with each other forces us to conclude that there are three separate phonemes, /p/, /t/, and /k/. The fact that [p] contrasts

with $[\overset{k}{p}]$ means that $[\overset{k}{p}]$ cannot belong to the /p/ phoneme; there remain three possible analyses for $[\overset{k}{p}]$.

(1) It belongs to neither /p/, /t/, nor /k/. It belongs to a phoneme by itself.

(2) It belongs to /t/.

(3) It belongs to /k/.

Analysis (1) is to be avoided if possible because we prefer not to set up more phonemes that we need to account for all the contrasts of the language. $[\overset{k}{p}]$ does not contrast with [t] or [k] since $[\overset{k}{p}]$ occurs only after rounded vowels while [t] and [k] occur only after unrounded vowels. We therefore reject analysis (1).

This leaves analyses (2) and (3). Given the choice of grouping $[\overset{k}{p}]$ with [t] or grouping it with [k] we do not hesitate to group it with [k] since phonetically it has more in common with [k] than with [t].

It may be asked why we were willing to assume that the final consonant was not affected by the initial consonants or by the length of the preceding vowel, but we were willing to assume it was affected by the preceding vowel's being rounded rather than unrounded. This is a reasonable objection, and the answer lies in considering the phonetic details carefully.

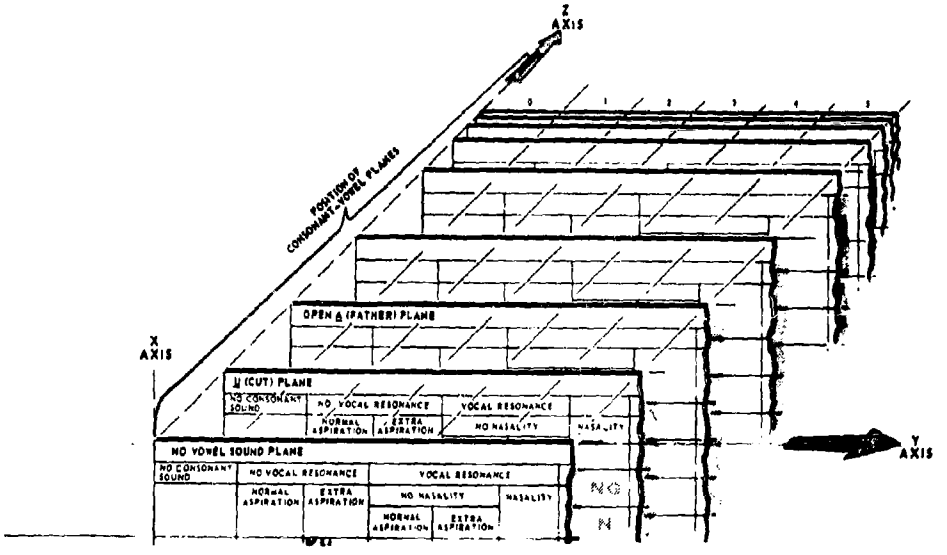
The initial consonant is not adjacent to the final consonant and although non-adjacent vowels sometimes influence each other directly (i.e. without changing any intervening sound), non-adjacent consonants rarely do so. This is a generalization which we believe holds true for all languages.

As for the final consonant being affected by the duration of the preceding vowel, this does happen in language, but the common effect is some change in the duration of the consonant. In the extreme cases, the consonant is dropped completely. We know of no cases, however, where the place of articulation of a consonant has changed audibly because the preceding vowel was phonemically long.

When we consider the effect of a rounded vowel or semivowel on an adjacent consonant, however, the situation is quite different. Such influences are common, and we know of one case in which a k followed by w became pp. This change took place in very early Greek, when the Proto-Indo-European word for "horse" became the Greek hippos, but remained almost unchanged in the Latin equus (pronounced ekwus).

We are citing this historical example, not to establish the origin of the Vietnamese [$\overset{k}{p}$], but to show that there is phonetic similarity between a p and a vowel or semivowel which involves lip-rounding. (We are here making the assumption that if one sound has been substituted for another in any language, there must be some point of phonetic similarity between the original sound and the substituted sound.) We are not concerned with how the [$\overset{k}{p}$] came into being, but how it functions in the language.

APPENDIX B - CHARTS OF THE CONSONANT CATEGORIES



Consonants shown as Reimann leaves of the plane on following pages.

STOPS AND NASALS

	Voiceless		Voiced		Voiced with nasal resonance
	Unaspirated	Aspirated	Unaspirated	Aspirated	
Guttural	k Fr. comme	$k^{\flat}$	g gu	$g^{\flat}$	ŋ sing
Palatal	ȼ	ȼ'	ɟ	ɟ'	n Sp. canon
Alveolar	t	$t^{\flat}$	d do	$d^{\flat}$	n sin
Dental	ṯ Fr. tiens	$t^{\flat}$	d Ger. du	$d^{\flat}$	n neun
Labiodental	p	$p^{\flat}$	b	$b^{\flat}$	m
Labial	p Fr. peut	$p^{\flat}$	b bear	$b^{\flat}$	m my

SIBILANTS

	Voiceless		Voiced		Voiced with nasal resonance
	Unaspirated	Aspirated	Unaspirated	Aspirated	
Guttural					
Palatal	ʃ		ʒ		
Alveolar	s		z		
Dental	s̪				
Labiodental					
Labial					

AFFRICATES

	Voiceless		Voiced		Voiced with nasal resonance
	Unaspirated	Aspirated	Unaspirated	Aspirated	
Guttural					
Palatal	č chin		ǰ join		
Alveolar					
Dental					
Labiodental					
Labial					

SPIRANTS

	Voiceless		Voiced		Voiced with nasal resonance
	Unaspirated	Aspirated	Unaspirated	Aspirated	
Guttural	ʕ				
Palatal					
Alveolar	þ ⱶ		ⱶ ð		
Dental	þ thin		ð this		
Labiodental	f fine		v vine		
Labial	ϕ				

LATERALS

	Voiceless		Voiced		Voiced with nasal resonance
	Unaspirated	Aspirated	Unaspirated	Aspirated	
Guttural			L > <u>bulk</u>		
Palatal			L <u>dull</u>		
Alveolar			l <u>lit</u>		
Dental			l < <u>well</u>		
Labiodental					
Labial					

APPENDIX C

In making a palatogram a plate is shaped so that it conforms to the contours of the roof of the mouth. This is then coated with a substance which changes appearance when it is touched by the tongue. After this has been fitted into the subject's mouth, he articulates the sound which is under investigation, and the plate is immediately removed. By examining the plate and determining just where its appearance has changed, we can establish which parts of the tongue make contact with the roof of the mouth during the articulation of the sound under study.

The paletograms in the texts are illustrations of what we believe the originals to be like, rather than original plates made by us.

APPENDIX D

Before giving Meyer's conclusions, we will list the tense and lax phone; this is as close as we can come to defining the terms.

The "tense" consonants include all voiceless consonants; the lax consonants include all voiced consonants except the liquids and nasals. The liquids and nasals are neither tense nor lax. The tense vowels include the vowels in the following words: wife, way, leaf, lose, lobe, and cloud. The lax vowels include the vowels of the following words: if, loss, less, gas, push, should, and bud. The list of tense vowels is incomplete; Meyer gives all his examples in the phonetic script used sixty years ago. Most of the words are recognizable, but a few are not. There are three tense vowels which we have not listed. Some of Meyer's conclusions apply to specific segments of the speech wave as we have divided it, but many do not. Those of his conclusions which refer to only one type of segment are given in the sections in which those segments are described. Those which group together two or more of our segments are as follows:

a. Consonant durations

(1) The duration of initial lax consonants in one- and two-syllable words is slightly shorter than the duration of initial tense consonants. The difference is greater for consonants in medial and final position.

(2) Apparently the duration of an initial consonant does not depend on the quality of the following vowel.

(3) Initial consonants in two-syllable words are slightly shorter than in one-syllable words; medial and final consonants in two-syllable words are significantly shorter than in one-syllable words.

(4) The duration of a final consonant is dependent on the quality of the preceding vowel; the higher the tongue position for the vowel, the longer the final consonant.

b. Vowel durations

(1) A lax vowel is shorter than a tense vowel.

(2) The higher the tongue-position, the shorter the vowel.

(3) A vowel before a tense final consonant is shorter than a vowel before a lax final consonant.

(4) A vowel before a stop is shorter than a vowel before a fricative.

(5) l, m, n, and ŋ tend to shorten the preceding vowel.

(6) The lengthening of a lax vowel under influence of the final consonant is slightly less for a naturally long vowel than for a naturally short one.

(7) The lengthening of a tense vowel under influence of the following consonant is considerably less for a naturally long vowel than for a naturally short one.

(8) The different vowel durations before different consonants cannot be explained as an attempt to keep the syllable duration or the rhythm constant.

(9) The duration of the stressed vowel in a two-syllable word is considerably shorter than the duration of the stressed vowel in a one-syllable word.

(10) A vowel before a tense medial consonant is shorter than a vowel before a lax medial consonant.

(11) The unstressed vowel of a two-syllable word is long.

(12) A vowel before a fricative (spirant, sibilant, [w], or [h] is longer than a vowel before a stop.

Meyer's conclusions are not directly applicable to the present model both because Meyer did not break the speech wave down enough for our purposes and because he analyzed British English. It is not enough to know that a vowel has been lengthened; we need to know which parts of the vowel have been affected.

APPENDIX E

i) Duration of Nasals

Using his tape recorder, Harrell reports that when the word mump is played backwards, the resulting combination is heard as mump (Harrell, 1958). He explains this by suggesting that an initial nasal is considerably shorter than a final nasal (at least in English), except in special circumstances.

In this case, however, mump is pronounced with an unreleased p in which the closure is not followed by a noise burst (as in a quick pronunciation of rump rather than oompah!). According to Harrell's hypothesis the only thing which makes an apparently final nasal as short as an initial one is in fact a following voiceless unreleased stop as in mump. In this special case the stop is articulated in the same place or "homorganic" with the nasal.

Since the p in such words as bump is frequently unreleased it may be necessary to instruct a machine for speech recognition that an apparently final nasal which is no longer than an initial nasal should be interpreted as a combination of nasal plus homorganic voiceless stop.

Meyer reports (Meyer, 1903) that a final nasal is approximately one and one half times as long as an initial nasal. Ilse Lehiste, however, reports data which appear to contradict this(Lehiste, 1960). Using comparison of spectrographs she performed experiments to discover characteristics of the sound wave that accompanied what is known as juncture. (The difference between an ice man and a nice man is that an ice man has a juncture after n, and a nice man has a

juncture before n.) According to Lehiste, the initial n of nice in the phrase a nice man is twice as long as the final n of an in the phrase an ice man. She suggests that this difference in the duration of n is an important cue for distinguishing these two phrases. That is, an initial n is recognised as being initial because it is longer. Meyer reports also that the n in an aim is shorter than the n in a name, but he does not comment on the fact that this seems to contradict his statement that final nasals are longer than initial ones.

This is a very complex problem and it requires further research. The answer may be that the word an in both these examples is completely unstressed while the words nice and name are both strongly stressed. The initial n's may have been lengthened because they are in the stressed syllable. This is the only hypothesis which occurs to us at present.

ii) Relative Duration of the Onglide, Steady-State and Offglide of Liquids and Semi-Vowels

Working with the Pattern Playback, Lisker attempted to make the machine produce intervocalic r, y, l, and w artificially. (Intervocalic means occuring between vowels; the actual sounds Lisker tried to reproduce were iri, ara, uru, iyi, aya, uyu, etc.)

In synthesizing artificial r, y, and w, Lisker discovered that the most recognizable sounds were created when he drew his spectrograph so that the onglide, steady-state, and offglide were of equal duration. In synthesizing intervocal l, however, the most natural sound occurred when the onglide and offglide were drawn slightly shorter than the steady-state (Lisker, 1957).

In another experiment work done at Haskins on synthesizing initial liquids and semi-vowels yielded the following information. The quality of the synthetic l was improved when the first-formant transition was made very short; r, y, and w were not adversely affected by having the first formant transition short. (O'Connor, Gerstman, Liberman, Delatre, and Cooper, 1957.)

iii) Duration of Spirants and Sibilants

Experimenting with the words use [yus] (noun) and use [yuz] (verb), Denes made tape recordings of this word-pair and established that the vowel preceding [z] was considerably longer than the vowel preceding [s]. (Denes, 1955)

The next step was to take the segment [s], shorten it, and put it after the vowel of [i u z]; and also to take the [z], lengthen it, and put it after the vowel of [i u s]. He reports that both combinations sounded like perfectly normal words. This would indicate that the duration of a sibilant is an important cue for identifying it as "voiced" or "voiceless".

Meyer's figures for sounds he defines as tense [f, þ, s] compared with sounds he defines as lax [v, ð, z] show that tense spirants and sibilants are longer than lax ones. They also show that [f, s, ʃ, v, and z] are slightly longer after lax vowels than after tense ones, while [þ] is considerably longer after lax vowels. He also reports that fricatives (i. e. spirants, sibilants, [h], and [w]) in general have a greater duration than stop closures.

iv) Duration of Stop Closures

Comparing spectrographs, Eli Fischer-Jørgensen reports that in Danish the closure of p, t, k is shorter than that of b, d, g (Fischer-

Jørgensen, 1954). Other work contradicts this, however. Leigh Lisker has recorded words with intervocalic b (as in rabid). He then cut out that section of tape which had the stop-gap on it. In its place he inserted a period of silence longer than the stop-gap of [b]. As a result the word was heard as having a [p]; rabid became rapid. Conversely if a tape of [p] has its stop-gap cut out and a period of silence shorter than the stop-gap of [p] is inserted, a [b] is heard; rapid is changed back to rabid. (Lisker, 1957).

It should be noted that this study was confined to stops between vowels. Whether similar results would be obtained for stops at the beginning or end of words is not known. The discrepancy between Lisker's findings and those of Fischer-Jørgensen also requires investigation. Since one study was made for Danish and one for English it is quite possible that both are valid.

Meyer reports that stop closures are shorter than the class of sounds which he calls fricatives (i. e. spirants, sibilants, [w], and [h]). He also reports that the closure of p, b is greatest, that of k, g next, and that of t, d least, and that the closures p, t, k show more variation in duration than any other class of speech sounds.

v) The Duration of Stop Bursts

There is some evidence that duration differs from stop to stop in other languages. Eli Fischer-Jørgensen (Fischer-Jørgensen, 1954), reporting on Danish stops, says g has a greater duration than d, which in turn has a greater duration than b. Similarly k is greater than t, which is greater than p. It should be noted that in Danish b, d, g are sometimes voiceless and p, t, k are sometimes

voiced. It may or may not be relevant to a study of English.

In producing stop consonants by use of machines Haskins Laboratories' reports indicate that the duration of the synthetic burst was .015 seconds (Cooper, Delattre, Liberman, Borst, and Gerstman, 1952).

vi) Duration Between the Stop Burst and the Beginning of the First Formant

Using the Pattern Playback Haskins Laboratories reports that a synthetic speech pattern which listeners perceive as voiced stop plus vowel, such as bah can be changed to one which listeners perceive as voiceless stop plus vowel, such as pa simply by cutting off the beginning of the first formants (Liberman, Delattre, and Cooper, 1958). The authors point out that since the voiceless stops in English are aspirated (pronounced with a half-heard h as in gat) while voiced stops are not (as in gat), the time-lag between the stop burst and the very beginning of the first formant is probably thought to be a period of aspiration; thus the stops are heard as voiced.

vii) Duration of Vowel Onglide, Offglide and Steady-State

a. Onglide

Again using their artificial speech machine, Haskins reports (Liberman, Delattre, Gerstman, and Cooper, 1956) that a pattern which is perceived as [bɛ], as in bet, changed to one perceived as [wɛ], as in wet, when the duration of the onglide or transition is increased. If the duration of the onglide is

increased still further, the result is [uɛ] (o͞o, ĕ). (It should be noted that the pattern which produced [bɛ] in this experiment consisted of formant transitions followed by steady-state. There was a voice-bar for the voiced stops, but no stop burst). The pattern for [gɛ], as in get, similarly yielded [yɛ], as in yet, and [iɛ], as in ē-ĕ, when the duration of the onglide (transition) was increased. [b], as in bet, was transformed to [w], as in wet, when the duration of the transition exceeded 40 milliseconds; [gɛ], as in get, became [yɛ], as in yet, at fifty to sixty milliseconds.

b. Steady-State

There is evidence that one important difference between the speech of Southerners and that of other Americans is the duration of the vowel steady-state. This evidence consists of photographs of spectrograms of regional speech in the book Visible Speech by Potter, Kopp, and Green (p. 11).

We have investigated some phonemes illustrating Southern speech. The most striking difference between the spectrograms of Southern speech and those of other regions of dialect is that the Southerner has much longer vowel nuclei, and explosive consonants. The duration of the steady-state portion of vowels and consonants could be accurately measured on a spectrogram. [illegible] longer steady state and shorter onglide [illegible].

It [illegible] did not reach the same steadi[illegible] may be due to the slowness of the [illegible] the normal steady-state level [illegible]. We may, therefore, [illegible] there is such a slow transition [illegible] to arrive at the steady state [illegible].

c. Offglide

Lehiste and Peterson report (Lehiste and Peterson 2, 1960) that American English vowels may be divided in to two groups using the criteria of the relative durations of steady-state and offglide. One group, which they call the tense vowels, consists of i, æ, a, ɔ, u ; the other group, which they call the lax vowels, consists of ɪ, ɛ, ə, ʊ. The tense vowels have an offglide approximately half as long as the steady-state; the lax vowels have an offglide more than one and one-fourth times as long as the steady-state. It should be noted that all the subjects who were used for this study spoke the same dialect.

APPENDIX F

The first method of reconstruction differs from the others in that it does not require any data except the language itself, considered at one point in time. For this reason it is called internal reconstruction. We have already cited the alternation of ə d , d, and t as past-tense markers in English. We have said that this is the result of two sound-changes, one in which ə d became d, and one in which d became t. We know that these two changes took place because we have eighteenth-century speech manuals which warn their readers not to omit the vowel of the past-tense suffix. Even if we did not have these manuals, however, we could still reconstruct part of the change by considering the nature of the alternation.

In any reconstruction we begin by assuming that one of the alternating sounds is the original one. If we cannot arrive at linguistically probable results by this approach, we then assume that all of the alternating sounds are innovations. In this English example, then, we will assume that ə d , d, or t was the original past-tense suffix. If we say that t was the original suffix, we must explain why t became d after the vowel e in laid , but remained unchanged after the same vowel in late . If t was the original suffix, laid and late were homonyms, and laid underwent a phonetic change while late remained unchanged. This conflicts with our basic principle that sound-change is regular, so we must reject the assumption that t was the original suffix.

This leaves us with the assumption that the original suffix was ə d or d. If we assume that it was d, we also assume that for at least a brief period of time, speakers of the language consistently pronounced the cluster td in the past tense of the verb taste. This is possible, but highly improbable. It is very doubtful that the original suffix was d.

This leaves us with the assumption that ə d was the original form. There are no difficulties involved in this assumption. The vowel could drop out quite easily after all sounds except d or t. The loss of the vowel would be a simplification of the articulatory movements, and many sound changes are simplifications of this type. After the vowel loss (or simultaneously with it) the d became t when it was next to a voiceless consonant. This sound change is also a simplification. The hypothesis that ə d was the original suffix involves our assuming only sound changes which are probable. The assumption that d was the original suffix involves our assuming a highly improbably state of affairs before the change. The assumption

that t was the original suffix involves our assuming a sound change which contradicts one of our basic postulates. We therefore conclude that əd was the original suffix.

The forms əd, d, and t stand in a special relationship to each other. They are different phonetic forms of the same meaningful element, the suffix for the past tense; which of them will appear depends on the phonetic shape of the last sound in the verb. This relationship is called morphophonemic alternation. There are other cases of this type of alternation in English; one of them is the plural suffix, which is əz after a sibilant or affricate (as in glasses), z after a vowel or voiced consonant (as in chairs), and s after voiceless consonant (as in books). Most morphophonemic alternations are entirely the result of sound changes; some are partly the result of changes by analogy. The difference between sound change and morphophonemic alternation is that sound change is process which takes place over a period of time, while morphophonemic alternation is a situation which exists at one time in the language. A machine for automatic speech recognition will need both a list of the morphophonemic alternations of that particular language and a list of sound changes which have taken place in any language. The list of morphophonemic alternations would make it unnecessary to make a separate statement about which alternate appears with each word. The list of sound changes will predict the normal sound variations of speech.

The second method of establishing what sound changes have taken place is to compare descriptions of the same language made at different times. For this comparison, we use only descriptions which are contemporary with the speech being described. In general the descriptions which we have were made for one of two reasons. The writer was either giving instructions on how to speak like a well-educated man or he was demonstrating the necessity for a spelling reform. Most of our descriptions of Latin, Greek, and eighteenth century English belong to the first category, while our best description of Old Icelandic belongs to the second. Both categories have particular drawbacks. The authors of pronunciation manuals sometimes make up rules which have never before existed, while the authors who recommend a spelling reform are primarily concerned with having a distinctive spelling for each phonetically distinct word. Their goal is not to record all phonetic differences, but all phonemic differences. Any statement which a writer makes about the pronunciation of his language must be carefully checked against the written records of that language, but these statements are nevertheless very valuable

for indicating what sound changes have taken place.

The third source of information about sound changes is written records. Although far better than nothing, these also have their drawbacks. One problem is that it is frequently difficult to determine what particular sound a given symbol or group of symbols is supposed to represent. We have very strong evidence for assuming that Indo-European had an s and that Old Icelandic had an r in place of that s in some phonetic environments. This means that s became r sometime between Proto-Indo-European and Old Icelandic. We have runic inscriptions from the period when this phonetic change was still taking place, but we do not know exactly what sounds the runes represent. There are three runes in question here. One occurs in those places where s did not become r; the second appears in those places in which s did not become r; the third appears in those places where we assume that there was an r in Indo-European which remained in Old Icelandic. If we knew what sound the rune for "s becoming r" represented, we would know the phonetic stages of this very common sound change, but all we can say definitely is that there were three separate sounds in Proto-Norse (the language of the Scandinavian runic inscriptions). Most phoneticians assume that when s becomes r the intermediate stage is z, and this seems probable, but we cannot prove it from written records.

The second drawback to written records is that spelling is usually standardized and does not necessarily reflect contemporary pronunciation. It would be extremely difficult, probably impossible, to reconstruct the pronunciation of Modern English using only written records other than dictionaries and speech manuals.

In analyzing written records, we can gain valuable information by paying careful attention to non-standard spellings (misspellings). A non-standard spelling is sometimes closer to the phonetic reality than the standard is. The incorrect nee reflects pronunciation more accurately than the correct knee does.

The fourth source of information about sound changes is the comparison of modern dialects of the same language. When two dialects show phonetic differences, it is obvious that one or both have undergone sound changes. One criterion for deciding what changes have taken place is simplicity. If two or more analyses seem equally probable phonetically, we assume the correctness of that one which involves the smallest number of changes.

The fifth method for discovering sound changes is comparing the earliest written records of two or more related languages for the purpose of reconstructing the parent language. The most extensive reconstruction of this type which has been done so far is the reconstruction of Proto-Indo-European by comparing written records of Greek, Latin, Sanskrit, Old Church Slavic, Hittite, Old Irish, and several other ancient languages. The phonetic accuracy of the reconstructed Proto-Indo-European forms is open to question. Some linguists say that these reconstructed elements should not be considered phonetic representations at all, but simply formulae for referring to the sound-correspondences of the later languages. According to this approach, "Proto-Indo-European p" is not the phonetic symbol p, but simply a formula for referring to that sound to Proto-Indo-European which became p in Greek, Latin, and Sanskrit, became f in Germanic, disappeared completely in Celtic, etc. Most linguists do not go quite this far, but there is general agreement that all Proto-Indo-European reconstructions should be critically analyzed in the light of phonetic probability. All methods of reconstructing sound change involve some possibility of inaccuracy, since there is no substitute for direct observation, and our reconstructions of Proto-Indo-European are especially likely to contain errors, since they are made from written records of languages which are now dead.

APPENDIX G

Martinet views linguistic evolution as something which is regulated by the continual conflict between man's expressive needs and his tendancy toward minimal mental and physical exertion. (Martinet, 1952). The "evolution" is the result of the changes in expressive needs which occur over a period of time. Martinet does not attempt to analyze in detail the changes in expressive needs. The principal effect of the expressive needs according to him is that the speaker strives for clarity. People strive to speak in such a manner that their enunciations can be understood without repetition. If they are not clear enough the first time, and the listeners ask for a repetition or an explanation, the speakers will be even more careful the second time. It is this process which results in some measure of uniformity in the speech of people. The striving for clarity is a clearcut factor which can prevent some sound changes.

The tendency to reduce mental and physical exertions to a minimum provides a more complicated problem in determining sound change; any change which reduces one type of exertion is likely to increase another. The extreme of articulatory simplicity would be to have two distinctive speech sounds (phonemes), one a vowel and one a consonant. All words in the language would consist of some arrangement of these two sounds in a series similar to that used by binary computers. The number of permitted phonetic variations of each phoneme would be extremely large, and this would save the speaker the trouble of having to articulate carefully. On the other hand, the words of this language would be excessively long, and any utterance would require a great deal of time and effort. The other extreme would be a language with as many distinctively different sounds as the human ear can perceive. Every sound would have to be articulated very carefully, but it would be possible to have very short words and utterances. Neither of these extreme cases exists in any natural language. In actual practice all languages require some precision, but none use more than a small fraction of all possible phonetic distinctions. This compromise requires less exertion from the users of the language than either of the extreme situations outlined above.

Exertion is further reduced by combining several distinct types of articulation to form a much larger number of phonemes. The efficiency of combining distinctive characteristics into phonemes is clear if we consider an imaginary language with four consonant phonemes, each of which has only one characteristic feature: (1) dental, (2) nasal, (3) voiceless, (4) spirant.

Although each phoneme has only one distinctive characteristic, each phone (speech-sound) must have many non-distinctive characteristics because the various articulatory organs must be in some position, and any position affects the quality of the resulting sound. Moreover, in the phonemic system under discussion, the characteristic which marks one phoneme must not occur with the allophones of any other phoneme as a non-distinctive characteristic. This is because, if a sound is uttered which has the distinctive characteristics of two phonemes, the listener would be unable to decide which phoneme the sound belongs to.

The speakers of this language, then, must be capable not only of the articulatory adjustments which produce a sound containing the distinctive characteristic, but also of the articulatory movements which will produce a sound lacking the distinctive characteristic. The speaker must be able to place his tongue in position for producing not only a dental sound, but also one with some other place of articulation. He must be able to produce not only a nasal sound, but one which is not nasal; not only a voiceless sound, but one which is not voiceless; not only a spirant, but also a sound with some other manner of articulation. It should be noted that most of these non-distinctive articulatory positions permit much more variation than the distinctive ones for this case. A dental sound must be made at the teeth, but a non-dental one may be made anywhere else in the mouth cavity. The non-spirant sound also has a large range of permissible variations. Even the non-voiceless sound permits some freedom, since in addition to normal voicing, the vocal flaps can also be placed in the position for laryngealization and trillization. The only articulatory position with no freedom is the non-nasal: the only way to produce a non-nasal sound is to have the velum completely closed.

The above described system is extremely inefficient. The speakers of this language are forced to discriminate between the presence and absence of four characteristics, but they have only four consonants. If they combined the distinctive characteristics, they could have many more consonant phonemes without having to learn to produce or to recognize any more distinctive articulations. In theory, they could have sixteen phonemes, but in practice the number would be less. A voiceless nasal spirant, whether dental or non-dental, would be difficult for the listeners to identify, and the speaker would frequently be asked to repeat his words. Since people do not like to make extra effort, we would not expect any language to have such a phoneme. There are other possible combinations of these characteristics such as voiceless nasals which are not optimally audible. Even if our hypothetical language fails to use any combinations of nasal with spirant or nasal with voiceless, it can still

make ten phonemes with the remaining combinations. This is far better than the original four, and the speakers and listeners are not required to make any new discriminations. Combining distinctive features to form phonemes is a very important method of reducing the exertions of those who use the language.

The term "distinctive features" as used herein, it should be noted, does not have exactly the same meaning as it does when it is used by Roman Jakobson. The concept of distinctive features originated in the Linguistic Circle of Prague in the 1930's. Both Jakobson and Martinet were members of this circle, but their ideas have developed along different lines since then. Jakobson maintains that there are only twelve distinctive features in all of the languages of the word, while Martinet argues that although the total number of distinctive features in any language is quite small, the total number of distinctive features which can be produced and recognized by human beings and which therefore may occur in some language is much larger than twelve. Moreover, Jakobson says that all oppositions are binary, while Martinet maintains that some oppositions are binary and some are not. Jakobson believes that the same distinctive feature can mark both the vowels and the consonants of the same language, while Martinet rejects this analysis because it would require too much precision on the part of the speaker. If vowels and consonants were marked by the same distinctive features, the speakers of the language would have to take care that the feature was not accidentally extended to an adjacent phone. If vowels and consonants are marked by different features, then the extension of a consonant-marking feature to an adjacent vowel does not interfere with communication.

Combining distinctive features to form phonemes saves exertions because it requires a smaller number of distinctive articulations, but it frequently requires greater precision for the articulation of some sounds than it does for others. Let us consider a language which has four different jaw positions that combine with distinctive tongue articulations to mark the vowels. If this language has four front vowels and four back vowels, the speaker will have to be more careful in articulating the back vowels than the front ones because although the jaw positions are the same, the tongue positions are clower together for the different back vowels than they are for the different front vowels. There is not as much vertical room in the back of the mouth as there is in the front, and therefore it takes much more precision to make three distinct levels in the back of the mouth. The speakers of the language may feel that the precision required involves too much effort, and they may make some change which reduces the number of back vowels to three.

The speakers of such a language may also try to simplify the tongue position for the back vowels. In so doing they may actually complicate the phonemes of the language.

The fact that non-integrated phonemes have more room for variation than integrated ones sometimes leads to the non-integrated becoming integrated. As the non-integrated phonemes vary, sooner or later some realizations of the phoneme may have a phonetic shape which makes them part of the already existing pattern. This happens when the language does not already have phonemes utilizing all practical combinations of distinctive features.

Let us consider the actual mechanism of sound change. We must always bear in mind that the principal difficulty in speech articulation is to produce just those sounds which are called for in a given context. A babbling baby can produce almost all the sounds of any human language, as well as some sounds which do not occur in any language, but he cannot control the production of these sounds. Great exertion is always easier than precision, and perfect control over all the articulatory movements is impossible. If tow speech sounds are acoustically identical, this is an accident, and it is an accident which very rarely happens.

In actual practice two phonetic realizations of the same phoneme may be quite dissimilar; the one point they have in common is that both lie within the normal range of the phoneme. The speakers aim for the "center of gravity," but they frequently go wide of the mark. If they go wide of the mark and fall too close to the "center of gravity" of another phoneme, the speaker has to stop and correct himself; but if they go equally wide of the mark in some direction where there is no other phoneme, this does not interfere with communication and may not even be noticed. In order to reduce the amount of precision required, languages leave a margin of safety, a "no man's land" between phonemes.

Martinet believes that many sound changes can be explained by the tendency of a language to maintain or increase its margin of safety. If we have three phonemes which are separated by equal margins of safety, and one begins to change and approach another, it may set off a chain reaction. In this particular case, the tendency to maintain the margin of safety is not as powerful as the force causing the change. The hypothetical case looks like this:

B A→ C

In this situation C must either change by moving further away from A or there will be a phonemic merger and some words which were formerly

distinct will become homonyms. If there is another place which is easily available to C or if the merger of A and C would break down the phonemic distinctions between a very large number of words which have the same general distribution (i. e., if the distinction of A and C has a high functional yield), C will change. If the distinction between A and C has a low functional yield, and if there is no place for C to go, there will be a merger.

As A moves away from B, this widens B's margin of safety with A. If B's margins of safety are narrow in all other directions, B may move towards the spot formerly occupied by A to widen its margins of safety with its neighbors. If these neighbors are also crowded, they may take advantage of the extra space, and there is a general shift which affects a large part of the system.

In the above discussion we have spoken as if A and B had an existence independent of the speakers of the language. This, of course, is not true. When we say that C moves in order to avoid a merger with A, we mean that the speakers of the language favor those variants of C which are a safe distance removed from the "center of gravity" of A. This results in a shift of the "center of gravity" of C.

If the merger of A and C would result in a few cases of intolerable homonymy, the existence of these few cases does not prevent merger, but one member of the homonymous pair drops out of the language, and it is replaced by another word of similar meaning. A sound-change which took place in southwestern France resulted in the words for 'cat' and 'rooster' being identical. In a farming community, it is necessary to have a distinct name for each domestic animal, so the old word for 'rooster' dropped out of the language, and was replaced by a word which had formerly meant 'pheasant'. When the speakers of a language are faced by situations of this type, they usually find a way out.

The fact that a contrast between two given phonemes has a high functional yield (many minimal pairs) means that a merger is extremely unlikely; but two phonemes will not necessarily merge when their opposition has a low functional yield (few minimal pairs). There are very few minimal pairs (pairs of words which are identical except that one member of the pair has phoneme A where the other has phoneme B) for the English phonemes /ð/ and /þ/ (or /θ/). The most commonly cited pair is thy and thigh, but it is extremely difficult to think of a context in which either of these is equally probable. Yet Martinet does not believe that these phonemes will merge, because the voiced feature is supplemented by accompanying variations in strength of articulation. In this manner ð is distinguished from θ, v from f, ʒ from s, ʒ̬ from š, ǰ from č, and b is separated from p, d from t, and g from k. The voiced-

voiceless opposition greatly helps to stabilize the consonant pattern of English, He cautions, however, that the possibility of phonetic change is not precluded by such an opposition, but concludes that a merger of sounds is less likely than if only one pair of consonants in the language were opposed (Martinet, 1952).

POSSIBLE APPLICATIONS OF MARTINET'S WORK TO OUR MODEL

One appect of the potential relevance of Martinet's theories to English may be considered in the light of the chart shown in Figure 17, which shows the stops, nasals, and spirants of English. The chart includes all the dimensions which were shown in the charts of our model, but we have arranged them differently in order to graph them on a single piece of paper.

According to Martinet, /f/ and /v/ can be considered to have the same phonemic place of articulation; he feels that a bilabial spirant would be very weak and difficult to recognize and therefore the speakers of the language have substituted a labiodental spirant. The distinctive feature, then, is labial articulation, which must involve the lower lip, while the other articulator is either the upper lip or the teeth. Martinet does not discuss whether the dental spirants can be considered to have the same place of articulation as the alveolar stops. The same argument which was used for /f/ and /v/ may apply here also. An alveolar spirant is not easy to identify, so it is quite reasonable that the speakers of the language should substitute a dental spirant. The distinctive feature would be apical (tongue-tip) articulation against either the teeth or the alveolar ridge. We assume, then, that all four spirants are well-integrated into the phonemic pattern. It is interesting to note, however, that there are no spirants corresponding to the guttural stops. This means that the speakers do not have to be as careful to make a complete stop closure for /k/ and /g/ as they must be for the labial and apical stops. It is possible that sometimes they substitute a spirant for a guttural stop. In his description of the phonetics of American English, C. K. Thomas (Thomas, 1947) gives one example of a word which formerly always had a [k] and now sometimes has a [ɣ]. ([ɣ] is a spirant with approximately the same place of articulation as the [k] of key.) The word technical is pronounced by some Americans with a spirant instead of a [k] before the [n].

Martinet's theories may also be relevant in considering the problem of /l/. There is only one lateral phoneme in English; this means that the sound pattern does not require that /l/ have a certain place of articulation or that it be voiced rather than voiceless. When we consider structure of the articulators and the mouth cavity, it

becomes obvious that there are physiological constraints on the production of a lateral. It is impossible to make a lateral with the lips or the teeth, and this immediately excludes bilabial and labiodental articulation. As far as voiceless laterals are concerned, we should bear in mind that a voiceless lateral is realtavely difficult for the listeners to recognize, although a few languages, such as Welsh, are reported to have voiceless lateral phonemes. We may expect some lateral articulations to be voiceless or partly voiceless, but not very many. The environments in which voiceless laterals are most likely to occur are after [p] and after [k], in such words as play and clay, and the voiceless lateral in these environments is apparently the result of a coarticulation. The tongue is in position for the lateral before the stop is released. Usually there is a short period of voiceless lateral followed by a short period of voiced lateral.

The place of articulation of the laterals is a far more complex problem because there are more variations involved. In an earlier report, we mentioned the fact that /l/ could show a great deal of phonetic variation, but at that time we believed that in a particular environment there would be little variation. We knew that the lateral most commonly occurring in bulk was quite different from the lateral most commonly occurring in lane, but we assumed that the lateral of bulk did not vary much from one pronunciation of the word to another. If Martinet's theories are correct, it may become necessary to question this assumption. It seems possible that the /l/ of bulk shows phonetic variation. The /l/ of this word probably has a guttural place of articulation more frequently than any other place, because it lies between a vowel and a /k/, both of which must be articulated precisely, and the lateral which requires the least exertion in this situation is a guttural one; but if the speaker feels that the guttural lateral is not distinct enough, he may shift the place of articulation forward in his mouth, either to the palatal or to the alveolar position. Our model must be specifically instructed to expect these variations.

In general we would expect a guttural lateral before a guttural and an alveolar lateral before an alveolar. We expect little variation in the place of articulation of a lateral before an alveolar, because alveolar articulation is the clearest, and in this environment it is also the easiest. In our acoustic research, we will investigate the problem of which lateral is most common before a labial.

There is probably little phonetic variation in an initial /l/ followed by a specific vowel, and since we treat initial consonants and following vowels as a unit, the phonetic difference between the /l/ of lay and the /l/ of low should pose no problems for the machine.

The final /l/'s of peel and pool have somewhat more freedom to vary, but since this variation will always be caused by the conflicts of simplicity and clarity the variation will be from the clearest articulation, which is alveolar, back towards the easiest, which will probably be palatal. In our acoustic research, we will seek to establish the easiest place of articulation for /l/ after different vowels.

Elsewhere in this report, we have given a list of sound-changes involving l. Judging by the nature of the changes, it seems clear that these l's are all dental or alveolar except where they are specifically described as having some other place of articulation. These rules will apply primarily to initial /l/'s, to final /l/'s after front vowels, and to /l/'s before alveolar consonants, since these are always alveolar.

We have discussed the laterals in detail because of all the sounds which we have included in our charts, these are the least "integrated."

The importance of the distinctive features is that when they are combined, the resulting sounds form a pattern. If a language combines the distinctive features of three places of articulation (labial, dental, and guttural), two vocal flap positions (voiced and voiceless) with a stop articulation, there are six possible phonemes, as follows:

	Labial	Dental	Guttural
Voiceless	p	t	k
Voiced	b	d	g

Each of these phonemes is integrated into the pattern; it is the product of combining a number of different distinctive features.

The above situation contrasts with the position of a phoneme such as English /l/. The distinctive feature of /l/ is lateral articulation, and no other phoneme has this feature. None of the stop consonants listed above has any feature which is unique. This unique characteristic of /l/ makes it completely non-integrated, while the stops are well integrated. English /l/ is free to vary phonetically, since neither the place of articulation nor the vocal flap adjustment are phonemic for it.

Martinet believes that a well-integrated phoneme is far less subject to individual change than a non-integrated one. In the pattern given above, the dental /t/ is not likely to become alveolar while everything else remains the same. If this happened, the number of places of articulation would be increased without increasing the number of phonemes. There would be an increase in exertion which would not be compensated for elsewhere, either by a reduction in exertion or by an increase in clarity. It is quite possible, however, that /t/ and /d/ might both change their place of articulation and become alveolar stops. This would involve no change in the total number of distinctive features.

One particular limitation to Martinet's work is inherent in the present development of phonemic theory itself. Researchers in the field of phonemics assume a direct relationship between distinctive features on one hand and resonance, place of articulation and manner of articulation on the other. Phonemic research from its early inception had defined phonemes in terms of resonance, place and manner of articulation. Researchers such as Martinet, Jakobson, and Halle described the phonemes of a language as having distinctive features that identify them from all other phonemes in a language. The distinctive features of a given phoneme, however, do not distinguish between possible allophones. Hence in Martinet's theories the distinctive features for k as in cook are identical with those of k as in kit , and by definition identical with the distinctive features of any other allophone of k. This poses a problem in transcribing acoustic data, since the phoneme k is described to include at least four different (but not distinctive) places of articulation in English speech, and the k of cook has a different place of articulation from the k of kit. Such a situation indicates very strongly that distinctive features do not characterize with necessary precision the resonances or place and manner of articulation of speech segments. This lack of precision makes it extremely difficult to interpret meaningfully the relationship between distinctive features of "phoneme" and the "center of gravity of its place of articulation."

An added problem in the use of distinctive features of descriptions of speech segments is the lack of precise relationship between these features and the acoustic characteristic of wave forms of speech. Such wavefforms, essential to the data included in our model, depend on the precise definition of resonance, place and manner of articulation; as indicated above, distinctive features lack such precise definitions.

Another important omission is that Martinet devotes little attention to the problem of the coloration which vowels give to preceding and following consonants. Our model includes a special subdivision for this subject particularly because articulation of a consonant is likely to be strongly influenced by the preparation which a speaker must make for his following vowels. The mechanics of sound change or variation in rapid speech are also likely to be influenced by what particular combinations of vowels and consonants occur together. In the case of sit down there is a definite transformation of the unvoiced dental stop, but in the case of look good there is less likely to be a transformation because the cluster occurs between identical vowel sounds. Such problems deserve further attention.

An additional problem is that Martinet's rules of sound change are at best only preferential; they have occasionally been observed to conflict with changes in the languages which are assumed to be relevant to English under the premises of the Ergodic theory and which in some cases have an observed relationship to sound changes actually taking place in rapid speech. Appendix H. 1 cites several examples of sound changes relevant to English that cross the boundaries of distinctive features. Among them are the gradual shift of a gutteral stop plus f to ff (as in big fire), the shift of a dental stop plus a gutteral stop to a gutteral stop (at camp), and the transformation of a dental stop before a labial stop to a labial stop (at bay). In all similar cases physical convenience in pronouncing the words may transcend the importance of distinctive features.

The inability of Martinet's theories to account for the sound changes discussed above argues that distinctive features at best have only limited relevance to our investigations. Dimensions of our model are constructed so as to account for many of the problems already discussed. By treating consonant vowels as one unit, we recognize the co-articulation of phones and avoid the question of different places of articulation for the same consonant caused by vowel coloration. By measuring duration we are able to analyze the process which can result in vowel coloration, changes in transition resulting in the articulation of a semi-vowel, difference in the target value of formants, and change in the characteristics of a sound through increased or lessened emphasis. By devising our new system of intensity measurement, already discussed, we are able to analyze in greater detail the influence of preceding and following sound units upon acoustic definition of a given phoneme while providing further information on the effects of emphasis.

Additional important information in our model is generated by specific investigation into the rules of euphonic combination. When a group of the allophones of any given phoneme combine in speech their characteristics may at times be modified to so great an extent that their acoustic signals may be identified as belonging to another distinct phoneme within the language; bet you becomes bechyou; at other times phonetic combinations result in total elimination of the signal for some phone classes, e. g. hot day becomes hoday. All such acoustic shifts form an essential and integrated basis for the definition of speech.

Any description of speech which does not consider such aspects may not be adequate to describe the acoustic characteristic of speech signals or to define a speech segmenting system suitable for actuating electronic machinery that can recognize speech by such characteristics.

Despite present limitations in data, however, the scope of our model suggests that distinctive features do not provide sufficiently full analysis of sound change for exclusive incorporation into our data.

APPENDIX H

RULES OF SOUND CHANGE AND EUPHONIC COMBINATION

H.1

CHANGES IN PLACE OF ARTICULATION

1. In guttural stops the following articulatory changes can occur:

   (a) A voiceless guttural stop + y becomes ss (Ionic Greek) or tt (Attic Greek); the second change may happen in cute kyut , but this is not likely.

   (b) A voiced guttural stop + y becomes zd (Greek); this does not happen in English.

   (c) A guttural stop + f becomes first a labial stop + f, then becomes ff (Latin). This may happen to such combinations as big fire in rapid speech.

   (d) The sounds k + t become tt (Vulgar Latin); this may happen in active but evidence seems to contradict it.

2. In dental stops the following changes occur (English has no dental stops, but these rules may also refer to alveolar stops):

   (a) A dental stop before a labial stop becomes a labial stop (Latin): this may happen in at bay.

   (b) A dental stop before a guttural stop becomes a guttural stop (Latin), possibly true of at camp.

   (c) A voiceless dental stop + y becomes s or ss (Greek); this does not happen in English.

   (d) A voiced dental stop + y becomes zd (Greek); this also does not happen in English.

(e) A voiceless dental stop + y becomes c in English; at you becomes a choo.

(f) A voiced dental stop + y becomes j; did you becomes di joo.

(g) A dental stop + m becomes mm (Latin). This may happen in English in the word atmosphere, but there is an important difference between the English example and the Latin. In English the difference between single and double consonant is not phonemic (not word-differentiating); single and double are probably in free variation in those positions where other languages have only double consonants.

(h) A dental stop + f becomes labial stop + f, then becomes ff (Latin); in English this change would not necessarily produce two f's for the same reason that atmosphere does not necessarily have two m's; at five is a possible example.

(i) A voiceless dental stop preceded by s becomes s + a voiceless lingual stop (Sanskrit); English has no lingual stops.

(j) A lingual stop + a dental stop becomes lingual stop + lingual stop (Sanskrit).

3. In dental nasals the following changes may happen:

(a) An n + guttural stop becomes ŋ (ng) + guttural stop (Greek); income nk m becomes ŋk m.

(b) An n + a labial stop becomes m + a labial stop (Greek); in bed becomes im bed.

(c) The sounds n + j become nj (Sanskrit), injure.

(d) The palatal stop + n becomes palatal stop + n (Sanskrit); possibly this occurs in such phrases as change now.

(e) An alveolar nasal n + m becomes m or mm in English; ten minutes becomes teminutes.

4. In dental sibilants the following changes can take place:

(a) An intervocalic sr becomes pr then becomes br (Latin); this probably doesn't occur in English.

(b) The sounds z + m become mm (Germanic); possibly this happens in is mine.

(c) A palatal + s becomes s (Sanskrit), possibly in church steeple.

5. In labial stops there are the following possible changes:

(a) A labial stop + a guttural stop becomes a double guttural stop (Latin); this could happen in up country, though the stop might not necessarily be double.

(b) The sounds p + t become tt (Vulgar Latin); this might happen in up to, although again the sound may not be germinate.

6. The following changes take place in labial nasals:

(a) A final m becomes n (Greek, Old English); possible in English, but not probable.

(b) The sounds m + d become nd (Germanic), possible in am down.

(c) The sounds m + s become s (Latin), possible in am so.

(d) A final m and non-labial stop become homorganic nasal + stop (Sanskrit), possible in am going.

7. The sound l becomes guttural or palatal in Latin and palatal in English:

   (a) when followed by a back vowel, low;

   (b) when followed by a consonant (except l), built;

   (c) when it occurs at the end of a word, ball.

APPENDIX III. II

CHANGES IN MANNER OF ARTICULATION

A. Changes Involving Stops

1. Changes from stop to spirant:

(a) In Germanic p becomes f everywhere except after a spirant or sibilant. In English, speakers may sometimes fail to make a complete stop closure, but we would expect the bilabial spirant [ɸ] rather than a labiodental spirant.

(b) In Germanic t becomes þ everywhere except after s or spirant. Again, in English a speaker might fail to make a complete closure, but we would expect an alveolar spirant rather than a labiodental one.

(c) In Germanic k becomes a voiceless guttural spirant everywhere except after a stop or spirant. In American English some speakers substitute a spirant for a stop in the word technical.

(d) In Modern Greek, all voiced unaspirated stops become spirants. We expect that this happens sometimes in Modern English; be may sometimes be pronounced with a [β] instead of a b.

2. Changes from stop to sibilant:

(a) In French k before front vowels becomes s. In Lithuanian Proto-Indo-European k becomes ʃ. If this happens in English, it affects words like kick, which may become sick or shick, but we do not expect this to happen.

(b) In Latin and Germanic, Proto-Indo-European tt becomes ss. If this happens in English, phrases like at two may become as sue, but we do not think that this occurs.

B. Changes Involving Spirants

1. Change from spirant to stop:

In Swedish [þ] becomes t. This happens sometimes in English. It affects words like thing.

2. Change from spirant to h:

In Spanish f becomes h. If this happens in English, it affects words like fine.

C. Changes Involving Hissing Sibilants

1. Non-combinatory changes:

(a) s becomes h in ancient Greek initially and between vowels. In English see may become he if it is weakly articulated.

(b) s becomes r between vowels in Latin and at the end of a word in Sanskrit if the next word begins with any voiced sound except r. (This change probably takes place in two steps. First s becomes z, then z becomes r.) We would not expect r as a variant of s in English, although s may become z, as indicated in Appendix H. III. (p. 153, rules 1 and 2).

(c) z becomes r in very early Old English and Old Norse. We are not sure this happens in English, even as a slip of the tongue. If it should happen, easy would be almost homonymous with eerie.

2. Combinatory changes:

(a) A dental stop plus s becomes ss in Greek and Latin. This probably occurs in English in phrases like at sea.

(b) sk becomes [ʃ] (the consonant of she) in Old English and Old High German. This may occur in English. If it does, it would make skip identical to ship.

(c) In Latin sr between vowels becomes [þr] (like the initial cluster of three). In standard American English, the cluster sr does not occur within a word, but þ is probably substituted for s sometimes in phrases like less rain.

(d) In Germanic zm becomes mm. This may happen in phrases like is more.

(e) In Latin sf becomes ff. This may happen in phrases like bus fare.

(f) In Modern English sy frequently becomes [ʃ], and zy becomes [ʒ]. This occurs in phrases like miss you and as you.

D. Changes Involving Laterals

1. Non-combinatory changes:

(a) Changes from stop to lateral articulation: in Latin d becomes l; In Sanskrit retroflexed d becomes retroflexed l. This may occur as a slip of the tongue in English: dot may become lot.

(b) Change from lateral to stop: In some French dialects ll becomes t in final position. There are two changes involved here — a change in manner of articulation and a change in resonances.

The change in manner of articulation probably took place first. We do not expect this change to take place in final position in English, since a final l usually does not have the same place of articulation as a t.

(c) Alternation between stop and lateral: In Greek d alternates with l. This means that either d becomes l or l becomes d. We have already suggested that dot may be mispronounced as lot; it is also possible that lot is mispronounced as dot.

(d) Change from lateral to sibilant: in Castilian Spanish l becomes [ʒ]. We do not expect this to happen in English.

(e) Changes from lateral to r and from r to lateral:

(1) In some Sanskrit dialects r becomes l. As a slip of the tongue, right may be pronounced as light.

(2) In some Sanskrit dialects l becomes r. As a slip of the tongue, light becomes right. Historically, this is what happened to the word colonel, which was once pronounced with an [l] between the o's. The [l] became [r] but the old spelling remained.

2. Combinatory changes:

(a) Loss of l:

(1) l drops out in unaccented syllable before t in Old Icelandic. This may occur in English in words like belt.

(2) l drops out before p in some dialects of American English. In these dialects help is pronounced [hɛp].

(b) Loss of another consonant in contact with l:

(1) In Greek ls becomes l and the preceding vowel is lengthened. If this occurs in English it may affect words like else, but as far as we know it does not happen.

(2) In Greek lw becomes l and the preceding vowel is lengthened. In English this may happen in such phrases as ill wind, but it is also possible that the w remains while the l drops.

(3) In Latin and in Old Icelandic initial wl becomes l. This cluster does not occur initially in English.

(4) In Old Icelandic nl becomes l and the preceding vowel is lengthened and nasalized. This probably occurs sometimes in words like inlay.

(5) In Modern English ly becomes a palatal l, as in the word million. This palatal l sometimes becomes y, so that million is pronounced [mɪyʌn]. This sound change commonly occurs in phrases like will you.

(c) Addition of a consonant to a cluster involving l:

(1) In Latin ml becomes mpl. This may occur in such English phrases as am late, although the rule which follows probably applies more frequently.

(2) In Greek ml becomes mbl. This probably occurs sometimes in the phrase am late.

(d) Complete assimilation of consonants in contact with l:

(1) In Greek and Latin nl becomes ll. English in late may

become il late, but English consonant durations vary freely, so that the phrase may also be pronounced i late.

(2) In Latin and Germanic ln becomes ll. This may happen in the English word illness.

(3) In Germanic, zl becomes ll. In English is late may become il late or i late. This would make the phrases in late and is late homonymous.

(4) In Latin ls becomes ll. If this happens in English, else, would become ell, but this probably doesn't occur.

(5) In Latin rl becomes ll. This may occur in phrases like here later.

(6) In Latin dl becomes ll. This probably occurs in phrases like bad link.

(7) In Sanskrit tl becomes ll. This probably occurs in phrases like at last.

(8) In Germanic _l becomes ll. This may occur in such phrases as with luck.

(9) In Greek, ly becomes ll. In English ly usually becomes a palatal l.

(e) Change in the manner of articulation of another consonant in contact with l:

In Latin v becomes b whenever it was preceded by l. This might occur in English phrases like full value.

APPENDIX H. III.

RESONANCES

A. 1. A voiceless consonant becomes voiced between vowels:

(a) when it occurs at the end of a word and the next word begins with a vowel (external combinations of Sanskrit), up at;

(b) when p, t, or k occur in the middle of a word (British Celtic); the difference between latter and ladder is in the vowel, not in the consonants;

(c) when s occurs between vowels; glassy becomes glazzy.

2. A voiceless consonant becomes voiced next to a voiced consonant:

(a) when a voiceless stop or s occurs before a voiced stop or spirant (Latin), up the becomes ub the;

(b) when p or k occur before d (Greek), back door becomes bag door;

(c) when an s occurs between a vowel and a voiced consonant (Latin), glass door becomes glaz door;

(d) when a final voiceless consonant precedes a word beginning with a voiced consonant (Sanskrit), Back Bay becomes Bag Bay.

3. A voiceless spirant or sibilant can become voiced when it occurs between voiced phones, except when the preceding vowel is accented (Proto-Germanic); in English we have the same situation in the phonetic difference between exert with gz and exercise with ks .

B. 1. A voiced consonant can become voiceless:

(a) if it occurs before a voiceless stop, spirant, or sibilant (Latin, Greek), big town may become bik town;

(b) when h occurs before w (some American dialects) where;

(c) when a voiced consonant occurs in the final position (Sanskrit — all voiced final stops; German — all voiced final stops, apirants and sibilants). In English, this may occur at the end of an utterance; where is the bag may become where is the bak. (The last word would still have a vowel like bag, rather than a vowel like back.)

APPENDIX H. IV

SOUND DROP OUTS

A. 1. In consonant clusters w will disappear phonetically:

(a) before l and r (Old Icelandic, Pre-Latin);

(b) after r, l, and n — while lengthening the preceding vowel in some dialects (Greek), bulwark may become bulark;

(c) medially before any consonant (Old Icelandic).

2. In consonant clusters y will drop out after n or r, while lengthening the preceding vowel (Greek), Bunyan may become Bunan.

3. In consonant clusters l will drop out:

(a) in an unaccented syllable before t (Old Icelandic), belt may possibly become bet if it is unaccented;

(b) before an s (Greek), Elsa may possibly become Esa, but we know of no instances where this has actually occurred.

4. In consonant clusters s disappears:

(a) between two consonants except two stops with the same place of articulation (Greek), pigsty might become pigty;

(b) after r or l, lengthening the preceding vowel (occasional example in Greek), hearse might become hɛ̄r . This change probably does not occur in English;

(c) after a vowel before a voiced consonant (Latin); glass door might become gla door, but the change probably doesn't occur in English.

5. Although English has no dental stops we have provisionally equated alveolar drop-outs with the dental stop drop-outs that occur in consonant clusters of other languages. Dental stop drop-outs occur:

   (a) before y after n (Germanic), plenty becomes pleny;

   (b) before s after n (Greek), landscape drops the d in some American dialects;

   (c) dental stop drop-outs also occur when an initial d precedes y (Latin). General American does not have this consonant cluster although some Southern dialects may, as in due. In this case the d probably does not drop, however;

   (d) In English a proven alveolar drop-out occurs when the stop comes between s and s; thus lasts frequently becomes lass;

   (e) an alveolar drop-out also occurs sometimes before l as when little becomes lil.

6. Guttural stops disappear:

   (a) after s before a consonant (Germanic), asked becomes ast;

   (b) after r or l and before s, t, m, or n (Latin), bulks (k possibly dropped);

   (c) when an initial g occurs before y (Latin); possibly this occurs in such English words as argue, but probably not.

7. Dental and guttural phones disappear before s plus a consonant (Germanic); this may occur in such English words as huckster.

8. In consonant clusters n disappears:

   (a) between a vowel and s (Latin), insert;

   (b) after a vowel before r or l, (Old Icelandic), inlay.

9. When it occurs in consonant clusters, m also disappears:

   (a) between a vowel and s (Latin), possibly also in English combinations such as come soon;

   (b) between a vowel and f (Latin, Old Icelandic); probably this occurs in such words as comfort.

10. The sound th disappears between consonants in English, as in fifths.

B. In positions other than consonant clusters the following sounds can drop out. Phonetic symbols suggest how this occurs in most cases as an aid to the reader.

   1. In final positions these sounds disappear:

      (a) m (Latin, Sanskrit, Germanic), m becomes voiceless nasal, then drops out completely; probably this does not occur in English;

      (b) n (Germanic, Old English), n becomes a voiceless nasal, then drops out completely; this may occur with man;

      (c) dental stops (Latin, Germanic); dental stop becomes a glottal stop, then drops out completely;

(d) the alveolar stop t , which can change to a glottal stop or drop completely under certain circumstances (Modern English), field or that boy, though probably only before a consonant. This loss of alveolar stops has been described by many phoneticians, among them C. K. Thomas in his book An Introduction to the Phonetics of American English (p. 40);

(e) s (Latin, when it occurs after u and the next word begins with an initial consonant); s becomes h, then drops out completely, or possibly s becomes z, then drops out completely; if this applies in English, loose connection would drop its s, but evidence seems to be against it.

2. Between vowels these sounds disappear:

(a) y (Old Icelandic), y becomes h, then drops out completely; crying may be reduced to one syllable;

(b) w (Greek; Latin when between like vowels and the second is unstressed); rowing may be reduced to one syllable;

(c) h (Greek), h drops out completely; in English is he becomes iz e.

3. The following initial sounds are likely to drop out:

(a) y (Old Icelandic), y becomes h, then drops out completely; we are not sure this happens in English;

(b) w (Greek); w drops out completely; we are also unsure that this happens in English;

(c) h (Modern Greek, Modern English when it is first in an unstressed syllable), h drops out completely; here's your book may possibly lose the h in here's.

4. The sound w will drop out particularly:

   (a) before a strongly rounded vowel (Old Icelandic); wool is a possible instance, though we do not presently believe this happens in English;

   (b) before o (Pre-Latin);

   (c) before u (Germanic);

   (d) after ō (Old Icelandic); probably this happens in English, as in low window.

5. Before a high front vowel y drops out (Old Icelandic); yield is an example, but this probably does not occur in English.

APPENDIX V

A. CHANGES INVOLVING PHONES WITH A SINGLE PLACE OF ARTICULATION

A. Changes Involving [y]:

Some of the rules listed here have been given elsewhere also. All rules concerning y are included here because they will help us fit y into the model.

1. Non-combinatory changes:

(a) Loss of y:

(1) In Old Icelandic, initial y is lost; it may become h before it drops completely. We doubt that this happens in English; if it does, it affects words like you.

(2) In Greek and Latin, y between vowels is lost. A similar variation in which y and an adjacent vowel are lost occurs in English words such as crying [kraɪyɪŋ], which becomes [kraɪŋ].

(b) Change of y:

(1) In Greek initial y becomes h. This probably does not happen in English. If it does, it affects words like you.

(2) In North Germanic, y between vowels becomes ggy. This does not happen in English.

(3) In Gothic, y between vowels becomes ddy. This does not happen in English.

(4) In Welsh, Proto-Indo-European y becomes ð after stressed e or i. We do not believe that this happens in English.

(5) In English h + y sometimes beoomes the voiceless guttural spirant[ç]. This occurs in the word hue.

2. Combinatory changes:

(a) Consonant clusters in which y is changed:

(1) A voiceless guttural or dental stop + y becomes ss in Ionic Greek and tt in Attic Greek. We do not believe that this happens in English. If it does, it affects words like cute.

(2) A voiced guttural or dental stop + y becomes zd in Greek. This does not happen in English.

(3) In Greek l + y becomes ll. In English this combination yields a palatal lateral. (See Appendix II. II. D - Changes Involving Laterals.)

(4) In Greek p + y becomes pt. This does not occur in English.

(b) Consonant clusters in which y remains:

(1) In Latin d + y becomes y in initial and medial position. In American English, except for some Southern dialects, the initial cluster dy-does not occur. In those dialects where due is pronounced dyu, this change may occur.

(2) In Latin g + y becomes y in initial and medial position. This may occur in American English. If so, it affects words like argue.

(3) In Sanskrit m + y becomes nasalized y + y. This may happen in English, but we doubt it; if it does happen, it affects phrases like come yet.

B. Changes Involving [r]

1. Non-combinatory changes:

(a) Loss of r:

In some dialects of American English r after vowels drops out. Many Southerners and New Englanders pronounce far without an r.

(b) Insertion of r:

In those American dialects which drop final r except when the following word starts with a vowel, an r is frequently inserted between a word ending with a vowel and another beginning with a vowel. A New Englander commonly pronounces deer that and idea that as [diə ðæt] and [aɪdiə ðæt], but he pronounces deer is and idea is as [dirɪz] and [aɪdirɪz].

(c) Change of r:

(1) In some Sanskrit dialects r becomes l. This probably happens in English; it affects words like right.

(2) In French r becomes z. As far as we know this does not happen in English.

(3) In Sanskrit r becomes s before an initial voiceless stop or sibilant. This probably does not happen in English.

2. Combinatory changes:

(a) Insertion of another consonant between a nasal and r:

(1) In Greek mr becomes mbr. This happens sometimes in English in phrases like come running.

(2) In Greek nr becomes ndr. This may happen in English. If it does, it affects phrases like in reference.

(b) Complete assimilation of another consonant to r:

(1) In Latin nr becomes rr. If this happens in English, it affects phrases like in reference.

(2) In Latin rs becomes rr. If this happens in English it affects phrases like for sale.

C. Changes Involving [h]

1. Non-combinatory changes:

Loss of h:

In Greek initial h and h between vowels drops out. This occurs in unstressed syllables in English. "Is he?" becomes [ɪzi].

2. Combinatory changes:

Cluster in which h and the other consonant are both modified:

In Modern English h + y sometimes becomes the voiceless guttural spirant [ç]. The word human is frequently pronounced [çumən].

D. Changes Involving [ʔ]

1. A final t before any initial bilabial consonant ([p], [b], [m], and [w]) is frequently replaced by [ʔ]. We have observed this in several different American dialects, including those of Virginia and Michigan; we believe it occurs in most dialects. Thus the phrase that one [ðæt wʌn] becomes [ðæʔ wʌn]; that boy [ðæt bɔɪ̯] becomes [ðæʔ bɔɪ̯]; atmosphere [ætməsfir] becomes [æʔməsfir].

2. A t before an l is normally replaced by [ʔ] in certain dialects, including that of New York City. Thus bottle [batl̩] becomes [baʔl̩].

3. A [ʔ] is sometimes inserted between a final vowel and an initial vowel. The ink [ðə ɪŋk] becomes [ðə ʔɪŋk].

APPENDIX H. V

B. CHANGES INVOLVING PHONES WITH TWO PLACES OF ARTICULATION

A. Changes Involving Labiovelar Stops

1. Changes involving retention of one place of articulation and loss of the other:

   (a) Retention of guttural (velar) articulation:

      (1) In Sanskrit k^w and g^w become k and g before a consonant and before a Proto-Indo-European back vowel. In Greek k^w and g^w become k and g before or after u.

      (2) In Latin k^w and g^w become k and g before a consonant or before u.

      (3) In Greek k^w and g^w become k and g before or after u.

      (4) In Germanic k^w and g^w become k and g before a rounded vowel. If this happens in English, it affects words like quote.

   (b) Retention of labial articulation:

      (1) In Oscan k^w and g^w become p and b in all environments. We doubt that this happens in English. If it does, it affects words like quit, quite, and quote.

      (2) In Latin g^w becomes w everywhere except after n or before a consonant or u. If this happens in English, it affects words like queen.

2. Changes involving shift in place of articulation:

(a) Shift to dental articulation:

In Greek k^w becomes t and g^w becomes d before a front vowel. We doubt that this happens in English. If it does, it affects words like quick and Gwen.

(b) Shift to palatal articulation:

In Sanskrit k^w becomes c and g^w becomes j before Proto-Indo-European front vowels. We doubt that this happens in English. If it does, it affects words like quit and Guam.

B. Changes Involving w

Some of the rules listed here have been given elsewhere also. All rules concerning w are included here because they will help us fit w into the model.

1. Non-combinatory changes:

(a) Loss of w:

(1) In Greek initial w is lost. This may not happen in English. If it does, it affects words like we.

(2) In Greek w between vowels is lost. This probably happens in English; it affects words like rowing, and reduces them to one syllable.

(3) In Old Icelandic w drops out after ǒ and before any strongly rounded vowel. In English this may affect phrases like low window.

(b) Change of w:

(1) In German and Latin w becomes v. This may happen in English; if it does, it affects words like we.

(2) In Welsh, initial w becomes gw. This also happens with Germanic loan words in French. This may happen sometimes in English, in words like with.

(3) In North Germanic w between vowels becomes ggw. This may happen sometimes in English, in phrases like bee wing.

2. Combinatory changes:

(a) Consonant clusters in which w is unchanged:

In Sanskrit m + w becomes nasalized w plus w. In English this probably happens in sentences like "Give him one."

(b) Consonant clusters in which w is lost:

(1) In Latin, dw between vowels becomes d. This probably does not happen in English. If it does, it affects phrases like add one.

(2) In Greek intervocalic lw becomes l, nw becomes n, and rw becomes r. In some dialects the preceding vowel is lengthened. This probably does not happen in English. If it does, it affects phrases like sell one, in one, and or one.

(c) Consonant clusters in which w is completely assimilated:

In Germanic nw becomes nn. This probably does not happen in English. If it does, it affects phrases like in one.

(d) Consonant clusters in which *w* and the other consonant are both changed:

(1) In Latin, initial *dw* becomes *b*. If this happens, it affects words like *dwell*.

(2) In Greek *kw* becomes *pp* and *gw* becomes *bb* before an *a*, an *o*, or a consonant; *kw* becomes *tt* and *gw* becomes *dd* before an *i* or an *e*; *kw* becomes *k* and *gw* becomes *g* before or after *u*. We do not believe that these rules apply to English.

APPENDIX H. VI

SANDHI RULES OF SANSKRIT AND THEIR APPLICATION TO ENGLISH

A. Phonetic Rules Relevant to English

1. When there are two or more consonants at the end of a word, the first is retained and the others dropped. This sometimes happens in English; act becomes ac and loft becomes lof.

2. Dental n coming after retroflex s or r, whether vocalic or consonantal, in the same word is changed to lingual ṇ. This change takes place even if a vowel, a semivowel, h, or any guttural or labial consonant comes between the r or retroflex s and dental n. This change does not take place if dental n ends a word. In American English, the n of internal has its place of articulation further back than an ordinary alveolar n. The place of articulation may be palatal.

3. Lingual r̥ followed by lingual r is dropped and the preceding vowel, if short, is made long. In American English this may happen with such words as or in phrases like or red.

4. Dental s following any vowel besides a or ā or following a guttural or a consonantal r becomes a retroflex ṣ. In American English the s of lease may have a palatal rather than an alveolar place of articulation.

5. When preceded by any stop consonant, h is changed to a voiced

aspirated stop having the same place of articulation as the preceding consonant. If this happens in English, it affects phrases like black hat.

B. Phonetic Rules Possibly Relevant to English

(We are not sure whether the sounds which we call "aspirated" in English are articulated in the same manner as the Sanskrit aspirates. If they are, then these Sanskrit rules may apply to English. We already know that some final voiceless stops in English, such as the k of back, are unaspirated. These rules may predict where.)

1. An aspirate stop or affricate is changed to a non-aspirate before another stop or before a sibilant; it stands unaltered only before a vowel, semivowel or nasal.
2. An aspirated stop becomes unaspirated in absolute final position (at the end of a sentence).

C. Historic and Analogic Rules

1. c or j is changed to k before voiceless consonants and g before any voiced consonant except a nasal or semivowel; this change also takes place when the consonants end a word, even before a nasal or semivowel.

(This rule represents an historic survival. At an earlier stage in the language, k and g become c and j in most phonetic environments, and remained unchanged in certain positions. The list of environments in which by sandhi rule c and j "become" k and g is really a list of the environments in which k and g did not become c and j. The

alternation between k and g and that between c and j are explained by the rules governing resonances in Appendix H, III.

This rule suggests that we should look for similar cases of historic survival in English. At present we have no examples.)

2. Final n followed by a dental, palatal, or lingual stop becomes ns. The stop remains unchanged. (We have already discussed this example in the main body of our text.)

3. The endings as and $\bar{a}s$ are governed by the following special rules:

   (a) When as is followed by a, it becomes o, and the following a is dropped.

   (b) When as is not followed by a, it becomes a and the resulting hiatus remains.

   (c) Before any voiced sound, $\bar{a}s$ loses its s and the resulting hiatus remains.

   (In Sanskrit most occurrences of final as were case-endings for nouns. These Sanskrit rules suggest the possibility that some English noun or verb endings, such as the possessive s, may have special rules governing their combination with the initial sounds of the following words. At present, however, we know of no such rules.)

4. The following rules governing final r appear to be analogic:

   (a) Before a pause, r becomes visarga.

   (b) Before a voiceless stop, r may become visarga.

   (c) Before a sibilant, r may become visarga.

   (We consider these rules to be analogic in origin for two reasons

first, we have no examples from any other language of r becoming an h like sound by regular phonetic change; and second, all the necessary conditions for an analogy were present in Sanskrit. Since s became r before a voiced sound, in many phonetic environments the s words had the same endings as the r words, and the s words were much more numerous. Under these circumstances, it is normal that the s rules should be extended to the r words.

We have not yet discovered any cases in English which show a similar alternation as a result of analogic change.)

5. Before a pause, s becomes visarga. If this happens in English, it affects words like space when they are in absolute final position.
6. Before any initial voiceless stop, final s may become visarga. If this happens in English, it affects phrases like space test.
7. Before an initial sibilant final s may become visarga. If this happens in English, it affects phrases like space shot.
8. Before an initial sibilant, final s may become a sibilant identical to the following one. If this happens in English, it affects phrases like space shot.
9. After any vowel except a or ā (in other words, after any vowel except one which has the sound quality of the first vowel of father, regardless of whether it is long or short), s becomes r before any voiced sound except r. If this happens in English, it affects phrases like space investigation.
10. Before an initial sibilant, final r may become a sibilant identical to the following one. If this happens in English, it affects phrases like more ships.

APPENDIX H. VII

A. Methods of Representing Euphonic Rules Symbolically

The rules shown below are in an invironmental form. The central phone or phones are those to which a particular transformation occurs. Those separated from the central ones by outward-facing brackets are the environment and represent the conditions under which a transformation will apply. Thus n] t [s represents the phone t in the environment "before s and after n." The result of a transformation is indicated by an arrow, so that a rule n] t [s ⟶ would mean that t drops between n and s, the blank space after the arrow meaning "no phone."

When a rule applies to some phone in more than one environment, this may be indicated by placing each environment on a separate line. For instance

$$\begin{matrix} n \\ \end{matrix}\Big] \quad t \quad \Big[\begin{matrix} s \\ s + t \end{matrix} \longrightarrow$$

would mean that t drops either when it is between n and s, or when it precedes an s followed by another t. The + notation means that one phone follows another, in this case t follows s on the second line of the rule.

It is often necessary to refer to whole classes of phones rather than to single ones. The symbols C and V refer to the classes of consonants and vowels. If particular characteristics are needed, these are placed in parentheses after the class or phone symbol, as C (v) for voiced consonants, n (D) for dental n, and so forth. (A complete

list of these abbreviations appears at the end of this section.) The rule following:

$$\begin{bmatrix} n \end{bmatrix} \quad (D) \quad \begin{bmatrix} s \\ s + \underline{C} \end{bmatrix} \longrightarrow$$

means that any dental phone is dropped between n and s or before s followed by any other consonant.

Since the dimensions of our model — manner of articulation, place of articulation, and resonances (aspiration and voicing) — play a special part, a three-position notation is often used. The three parts, separated by commas, represent the three dimensions, respectively. Thus (Af, Ld, -a+v) represents the phone v which is articulated as an affricate, is articulated in the labiodental position, is unaspirated and voiced. (See list of abbreviations.) This notation permits us several conveniences. We may leave a position blank to indicate that any "value" in that dimension is valid in the rule desired. We may omit the letters a and v for aspiration and voicing, so that v might be written (Af, Ld, -+), and any unaspirated, voiced phone could be written (, , -+). We may place one symbol above another to indicate several choices for one dimension, so that $(\begin{matrix} Af \\ Sp \\ s \end{matrix}, \begin{matrix} D \\ A, \end{matrix})$ would mean any dental or alveolar affricate, spirant, or sibilant. (See abbreviations.)

Using either form of the parenthesis notation, we may further economize on symbols by indicating several phones with shared characteristics using a single parenthesis and placing the symbols one above the other, as $\begin{matrix} n \\ l \end{matrix}$ (D) for dental n or dental l.

When an environment allows for several phones to precede the same following set of phones, the shared phones may be written on

the same line with commas intervening, as n, s] t [s for n t s or s t s, or $^{n,\,s}$] t [s, (St) / s + C meaning any of the five environments n] [s , s] [s , n] [(St) , s] [(St) ,] [s + C. For convenience, when no preceding or following environment is specified, the appropriate bracket may be omitted. No confusion should result, as the brackets are always directed so as to contain the environmental phones rather than the central phones. The symbols t [s / V then mean t before s or a vowel.

A few special symbols are used. Superscripts $^{+}$, *, $^{''}$ mean respectively "lengthened", "high stress", "low stress", so that V^{+} is a lengthened vowel, $(v)^{''}$ is any unstressed voiced phone. The letters x, y, z inside parentheses are reserved as variables. The rule (St) [(St, x, yz) ⟶ (St, x, yz) means that a stop coming before another stop takes the same place of articulation, aspiration, and voicing as the following stop. A rule (St, x, ·) [(St, x,) ⟶ means that a stop preceding one with the same place of articulation is dropped. Subscripts are used to indicate identical phones, so that V_1] w [V_1'' means w between two identical consonants, the second being unstressed. * and ** mean word-break and end of sentence.

The following symbols are used in the rules:

1) Manner of Articulation

St	Stop	Lg	Lingual
N	Nasal (nasalized)	R	Retroflex
Af	Affricate	V	Visarga
L	Lateral	Cs	Consonantal
S	Sibilant		
Ch	(Characteristics yet to be determined)		

2) Place of articulation

Gl	Glottal	D	Dental
G	Guttural	Ld	Labiodental
P	Palatal	B	Bilabial
A	Alveolar		

3) Resonances

+a	Aspirated	-a	Unaspirated
+v	Voiced	-v	Unvoiced

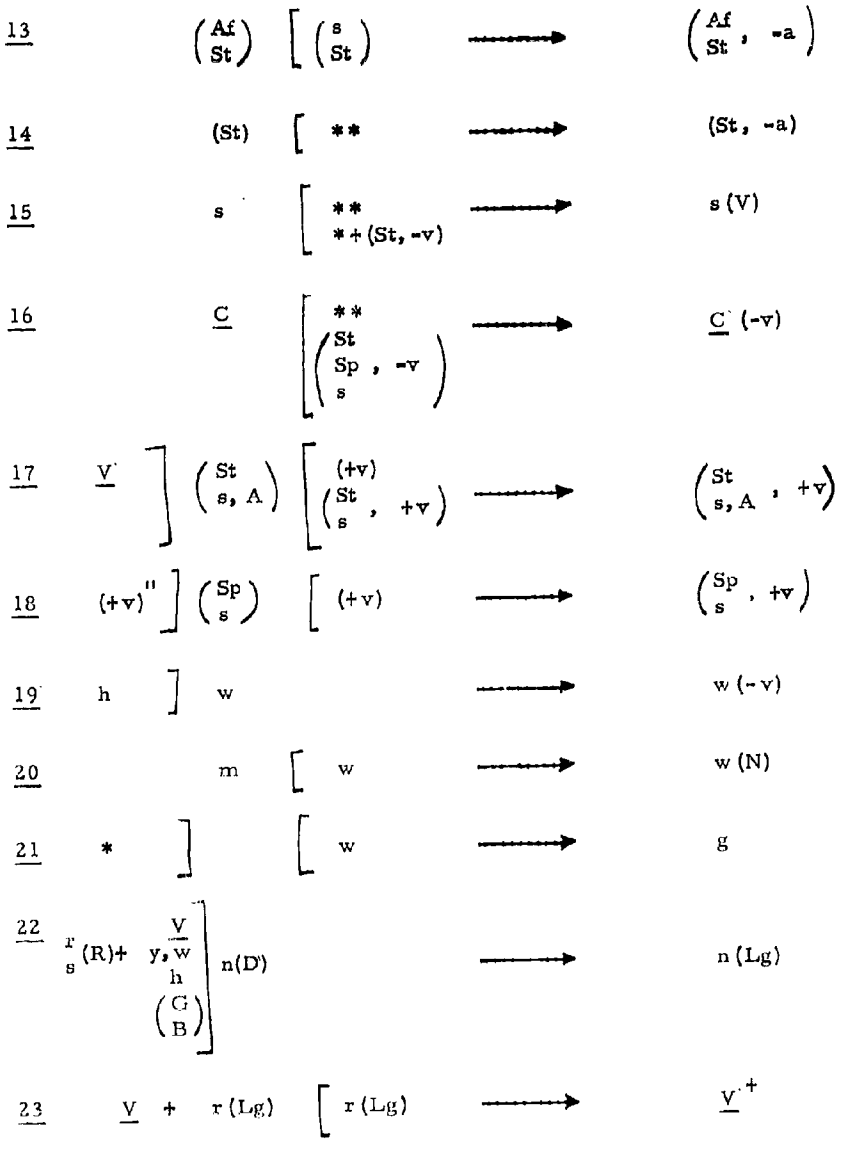
4) Others

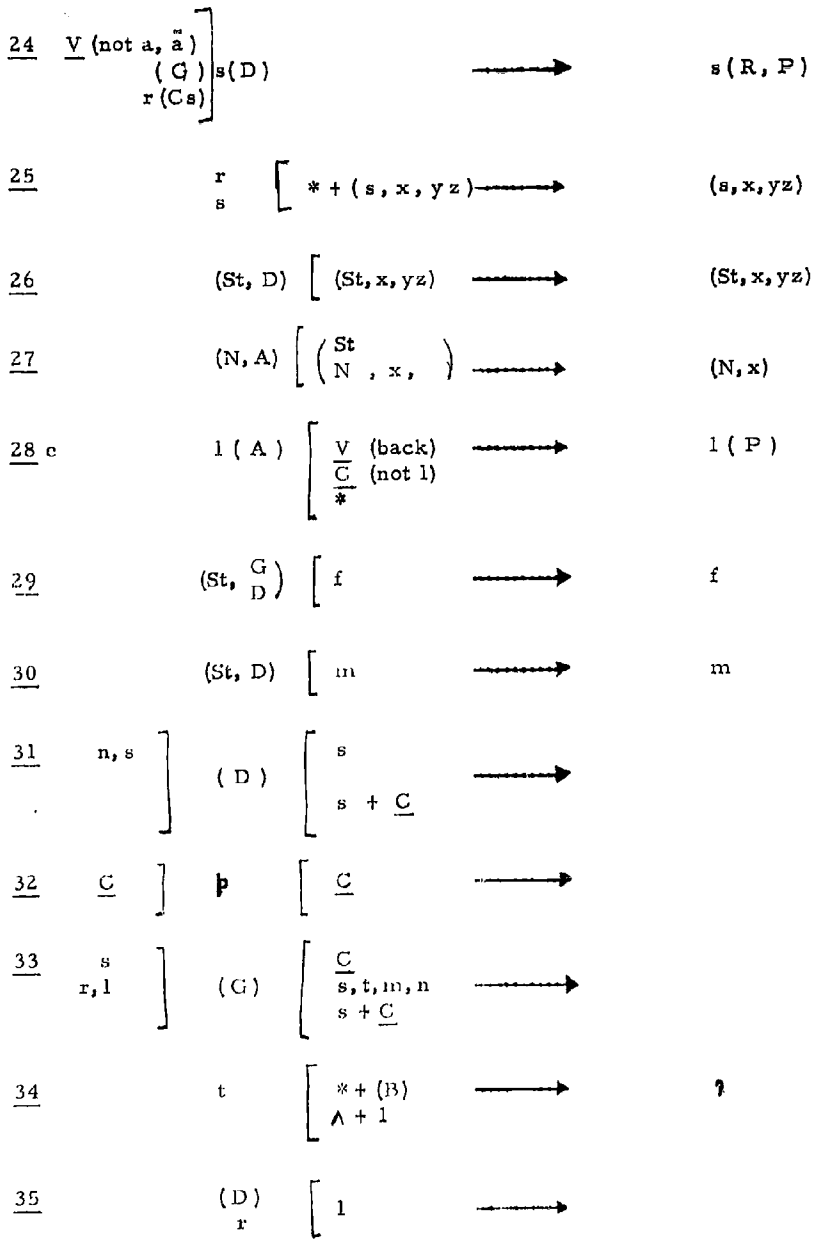
C	Consonant	"	Unstressed
V	Vowel	x, y, z	Variable characteristics
*	Pause	1	Identical phone
**	Final Pause		
+	Lengthened		
'	Stressed		

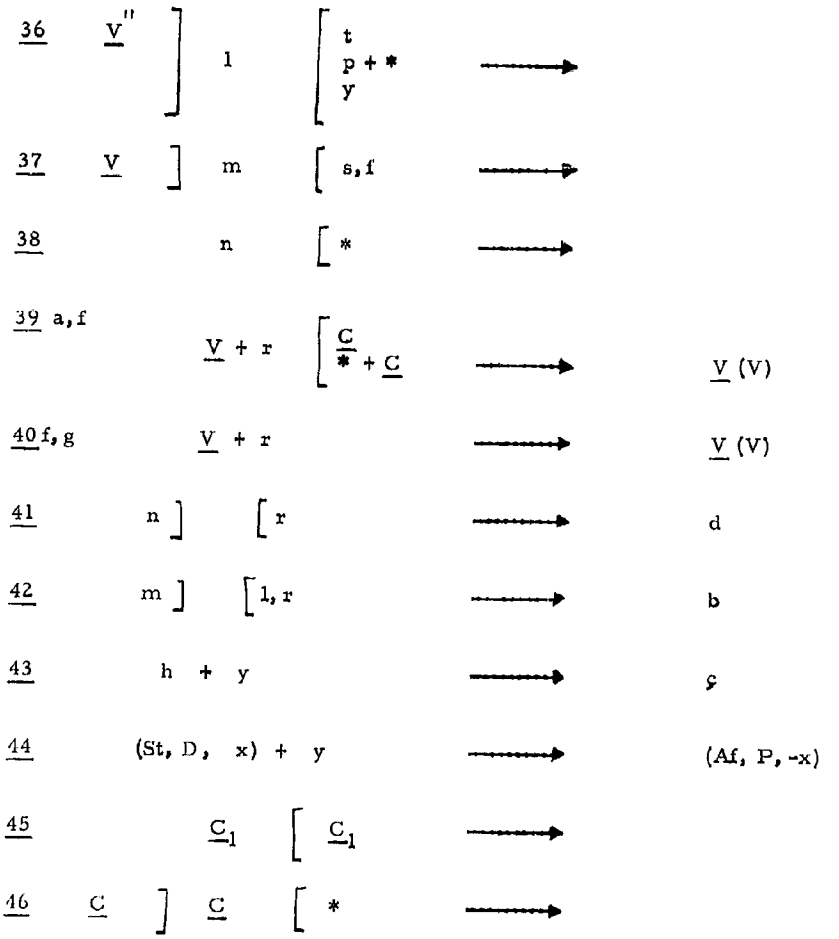
B. Symbolic Rules for Euphonic Combination

(Lower case letters following rule numbers refer to notes at the end of the list.)

1	V]		C	[* + V * + (+v)	⟶	C (+v)
2 a	C]		*	[h, V	⟶	
3 b	V]		*	[V	⟶	r
4 a	V (ch)]		*	[V (ch)	⟶	?
5 a	V]		*	[V	⟶	
6 c	V]		h y	[V	⟶	
7 c	$\bar{o}$ V_1]		w	[V_1''	⟶	
8	(St, x,)]		h		⟶	(St, x, ++)
9			h	[V''	⟶	
10 c	V (ch)]			[V (ch)	⟶	y
11 c	V (ch)]			[V (ch)	⟶	w
12 d	V (ch)		+	V (ch)	⟶	V (ch)


13
14
15
16
17
18
19
20
21
22
23
(Af St) [(s St) ⟶ (Af St , -a)
(St) [** ⟶ (St, -a)
s [** ** + (St, -v) ⟶ s (V)
C [** (St Sp , -v s) ⟶ C (-v)
V] (St s, A) [(+v) (St s , +v) ⟶ (St s, A , +v)
(+v)''] (Sp s) [(+v) ⟶ (Sp s , +v)
h] w ⟶ w (-v)
m [w ⟶ w (N)
*] [w ⟶ g
r s (R) + y, V w h (G B)] n(D') ⟶ n (Lg)
V + r (Lg) [r (Lg) ⟶ V +


24 V (not a, ā) (G) r(Cs) s(D) → s(R, P)
25 r s [* + (s, x, yz) → (s, x, yz)
26 (St, D) [(St, x, yz) → (St, x, yz)
27 (N, A) [(St N , x,) → (N, x)
28 c 1 (A) [V (back) C (not 1) * → 1 (P)
29 (St, G D) [f → f
30 (St, D) [m → m
31 n, s] (D) [s s + C →
32 C] þ [C →
33 s r, l] (G) [C s, t, m, n s + C →
34 t [* + (B) ∧ + 1 → ?
35 (D) r [1 →


36 V''] l [t p + * y
37 V] m [s, f
38 n [*
39 a, f V + r [C * + C
40 f, g V + r
41 n] [r
42 m] [l, r
43 h + y
44 (St, D, x) + y
45 C1 [C1
46 C] C [*
V (V)
V (V)
d
b
ç
(Af, P, -x)

NOTES:

a. Part of a pause-dropping scheme not yet fully developed. We know pause does not drop before bilabials.

b. New England Dialect only.

c. y and w may add or subtract between vowels, but the exact conditions are not well known. Acoustic data, some of which is included in this report may soon be able to further clarify these conditions.

d. Although we do not yet know the conditions for vowel coalescence, this rule is a vital part of the scheme for handling vowels.

e. The condition (not 1) is spurious since we regard a "double" of any phoneme other than a stop as a mere lengthening. (See also rule 45.)

f. Some dialects omit the visarga.

g. Southern dialect only.

C. Relation of Present Symbolic Rules to Rules of Euphonic Combination Previously Developed

The list following shows how the data presented in the April and May reports were incorporated into the mathematical formulation of rules for articulation. Rule numbers correspond to those in part B. of this Appendix A plus sign indicates that the combined effect of several rules must be used to achieve the result of the verbal description. Abbreviations used are:

na	Not applicable to English
nt	Not true for English
p	Partially used in . . . (rule number)
c	Contradicts . . . (rule number)
d	Doubtful validity
al	May be added at a later time.

Appendix H. I.

1. (a)	nt	2. (e)	44	3. (c)	al	6. (a)	nt
(b)	nt	(f)	44	(d)	al	(b)	al
(c)	29	(g)	30	(e)	27	(c)	al
(d)	nt	(h)	29	4. (a)	nt	(d)	al
2. (a)	26	(i)	na	(b)	d	7. (a)	28
(b)	26	(j)	na	(c)	al	(b)	28
(c)	nt	3. (a)	27	5. (a)	al	(c)	28
(d)	nt	(b)	27	(b)	al		

Appendix H. III

A. 1. (a)	1	2. (a)	17	2. (d)	1	(b)	19
(b)	17	(b)	17	3.	18	(c)	16
(c)	17	(c)	17	B. 1. (a)	16		

Appendix H, IV

A. 1. (a)	al	5. (b)	38	9. (a)	37	2. (c)	6
(b)	al	(c)	nt	(b)	37	3. (a)	d
(c)	al	(d)	38	10.	32	(b)	d
2.	al	(e)	c 17, 34	B. 1. (a)	nt	(c)	d
3. (a)	36	6. (a)	33	(b)	38	4. (a)	d
(b)	nt	(b)	33	(c)	p 34	(b)	d
4. (a)	c 33	(c)	nt	(d)	p 34	(c)	d
(b)	nt	7.	38	(e)	nt	(d)	7
(c)	nt	8. (a)	al	2. (a)	6	5.	d
5. (a)	c 44	(b)	al	(b)	7		

Appendix II. VI

A. 1.	16	B. 1.	13	(b)	na	5.	15
2.	22	2.	14	(c)	na	6.	15
3.	23	C. 1.	d	4. (a)	al	7.	c 25
4.	24	2.	na	(b)	al	8.	25
5.	8	3. (a)	na	(c)	al	9.	15
						10.	25

Appendix H. II

A.1.(a) d
(b) d
(c) d
(d) al
2.(a) nt
(b) nt
B.1. d
2. d
C.1.(a) d
(b) p 17
(c) nt
2.(a) al
(b) d
(c) d
(d) al
(e) al
(f) al
D.1.(a) d
(b) nt
(c) d
(d) nt
(e)(1) d
(e)(2) d
2.(a)(1) 36
(2) p 36
(b)(1) nt
(2) d
(3) na
(4) al
(5) 36
(c)(1) c 42
(2) 42
(d)(1) d
(d)(2) d
(3) al
(4) d
(5) 35
(6) 35
(7) 35
(8) 35
(9) 36
(e) d

Appendix H. V. A

A.1.(a)(1) d
(2) 6 & 12
(b)(1) d
(2) nt
(3) nt
(4) nt
(5) 43
2.(a)(1) nt
(2) nt
(3) 36
(4) nt
(b)(1) d
(2) d
(3) d
B.1.(a) 41
(b) 3
(c)(1) al
(2) nt
(3) nt
2.(a)(1) 42
(2) 41
(b)(1) c 41
(2) c 25
C.1. p 19, p 6
2. 43
D.1. 34
2. 34
3. 4

Appendix H. V. B.

A.1.(a)(1) na
(2) na
(3) na
(4) na
(b)(1) na
(2) c 21
2.(a) d
(b) d
B.1.(a)(1) d, c 21
(2) p 7
(3) p 7
(b)(1) d
(2) 21
(3) d
2.(a) 20
(b)(1) d
(2) d
(c) d
(d)(1) d
(2) d

APPENDIX H. VIII

RULES OF SOUND SHIFT DERIVED FROM MARTINET'S THEORY OF MINIMUM EXERTION IN ARTICULATION FOR POSSIBLE USE IN OUR MODEL

1. Before or after a dental, an alveolar may become a dental, as in health and width.
2. Before or after a labial or labiodental, an alveolar (except l) may become a dental, as in apt and at peace.
3. Before or after a guttural, an alveolar may become a palatal in words like books and act.
4. Before or after a labial or a labiodental, a palatal may become alveolar. This occurs in phrases like ash bin.
5. Before or after a labiodental or an alveolar, a palatal hushing sibilant may become alveolar. This occurs in phrases like red shoes.
6. l is basically an alveolar consonant, but before or after a back vowel, a palatal consonant, or a guttural consonant, it normally takes the place of articulation of the adjacent phone; it may alternately have its place of articulation at the alveolar ridge, or between the alveolar ridge and its normal place of articulation if the speaker is concerned about clarity.
7. Before a labiodental, a labial may become labiodental. This occurs in phrases like am fine and clip four.
8. After an r, an affricate may become an alveolar. In church, the second [č] is farther forward than the first.

APPENDIX H.IX
FURTHER RULES OF SOUND CHANGE

(This Appendix includes rules which have not previously been discussed or listed, rules which have been discussed in the text, but which have not been listed in previous sections of this Appendix, and rules which have been revised to agree with acoustic data studied.)

1. A final consonant of one word may be attached to an initial phone of the following word, as in phrases like make-up and made of. (See discussion on page 58 of the text of this report.)

2. When a nasal is followed by a voiceless sibilant or spirant, a voiceless stop may be inserted between them; mince may be pronounced mints. (This rule has not been discussed previously.)

   On page 40 of the text, we noted that sometimes "the t almost disappears from rents." This example seems to contradict this rule, since according to the rule, a t would probably be inserted in such a phonetic context. Truby has shown (Truby, 1959, p. 206) that the words prince and prints are homonyms; sometimes there is a stop-gap present and sometimes not, but prince has a stop-gap as often as prints. The same thing is true of lens and lends.

3. A voiced consonant in a voiced environment may become voiceless and sometimes aspirated; this may affect words like rouge and begin. This rule has not been discussed previously.)

4. A vowel may be replaced by a visarga vowel, as in the New England pronunciation of words like car and yard. (See discussions on page 58 of the text and in Appendix I. In Appendix H. VI (p. 171) visarga is referred to as "analogic" in nature -- the acoustic data definitely indicates that such sound changes occur in English.)

5. Final d may drop out after n, as in words like land. (This rule has not been discussed previously.)

6. A glottal stop may occur before any initial vowel, as in words like one, at, and among. (This rule was originally included in Appendix H. V. -page 164, but it is being revised here to conform to evidence found in the data studied.)

7. An alveolar stop between two consonants may drop out. (In Appendix H. IV, we said that a t may drop between two s's. We are here enlarging the scope of this rule.)

8. An alveolar stop before an affricate may drop out. (This rule has not been discussed previously).

9. When two identical consonants come together, they form one long consonant, that is, a consonant in which one part is considerably longer than normal. (In the case of stop consonants this long part is the stop gap; in the case of spirants and sibilants it is the fricative portion.) This long consonant may be shortened to the normal length of a single consonant.

This rule applies to any combination of identical consonants, re-

gardless of whether both these consonants are normal in the particular words involved or whether one is the product of another rule of euphonic combination. (This was discussed in Section 3 - page 52 of the test - but it was not included in the rules.)

10. In some New England dialects an r after a vowel is dropped except when r stands at the end of a word and the next word begins with a vowel. (This was discussed on page 57 of the text but was not included as a rule.)

11. In some New England dialects the r may be inserted between a word ending with a vowel and a word beginning with a vowel. (This was discussed on page 57 of the text, but was not included as a rule.)

APPENDIX I

VISARGA VOWELS

Visarga is a class of vowels defined by Sanskrit phoneticians as sounds that are co-articulations of vowels with h or aspiration.

This class of vowels is not mentioned among most European vowel sounds, and it has not been referred to in acoustic phonetic studies. The commonly used methods of phonemic analysis of languages, described in our previous report, probably do not yield such phonemes in carefully articulated speech of the languages referred to.

However, if the freedoms in the articulation of speech make it possible to generate such sounds, then they could occasionally occur in continuous speech of one or more of these languages, even if the presence of visarga is not formally recognized for these. For this reason, the generation of visarga and the expected acoustic characteristics of its waveforms are considered next.

For generation of visarga, the position of cheeks, tongue and lips are the same as those for the vowel of its form. The only difference between a visarga vowel and a normal vowel is that there is a steady

stream of air flow from the vocal cords for the former as opposed to significant periodic interruption of such a flow for the latter. This is illustrated in Figure 75.

Since the output of the vocal flaps passes through the mouth cavity that represents situations which are similar for both types of vowels, the formant frequency levels are expected to be about the same for both these types of vowels. However, since there is some difference in the spectral characteristics of the sounds of these two types of vowels; there should be some difference in the formant frequency levels, also.

The presence of flow of air, as in the articulation of h, should produce additional frictional energy in the case of visarga vowels, whereas such energy is absent in normal vowel articulation.

With the preceding discussion of visarga vowels it is worth considering the possibility of their occurrence in the English language.

When one considers an expression ending in a sigh or an exclamation indicating a relief, there is a distinct possibility of generation of visarga vowels.

Another possible occurrence of visarga vowels could be the vowels in the Boston accent that precede an ending r, such as a in car; but this aspect needs to be substantiated.

Since the freedoms in the methods of speech generation indicate the possibility of production of visarga vowels, and since the rules of

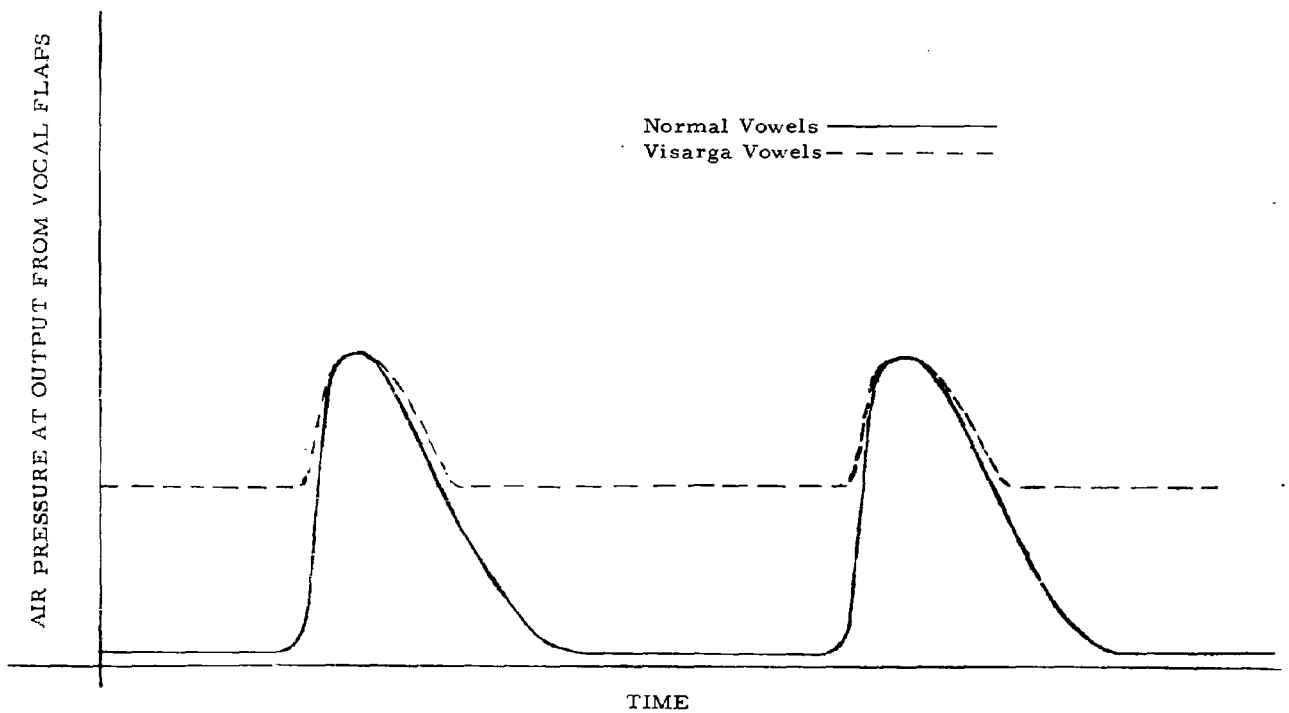

AIR PRESSURE AT OUTPUT FROM VOCAL FLAPS
Normal Vowels
Visarga Vowels
TIME

Figure 75

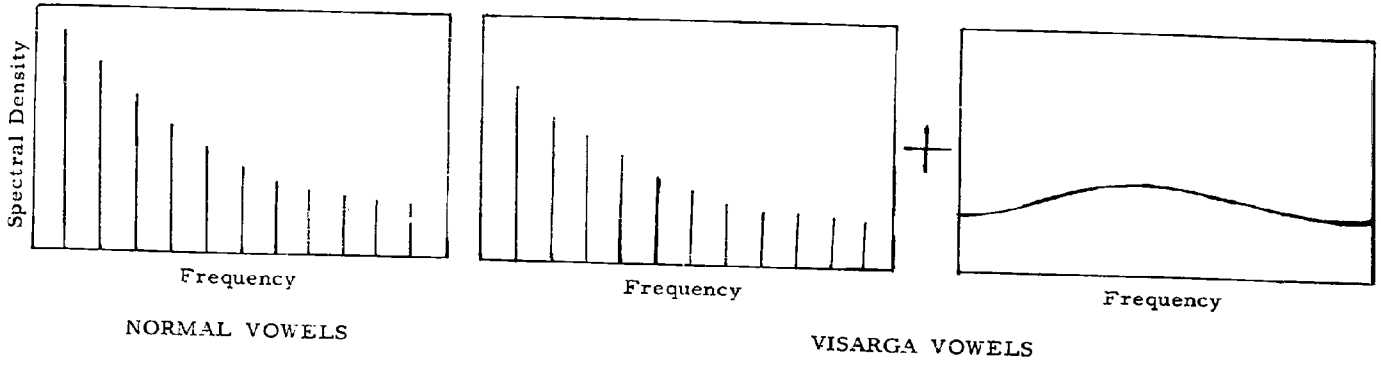

Spectral Density
Frequency
Frequency
Frequency
NORMAL VOWELS
VISARGA VOWELS

Figure 76

euphonic combination indicate the possibility of their occurrence in English conversation, these vowels are added along the vowel dimension of the model.

For conditions of production of visarga, described above, the vocal flaps can be considered to be partly open at all times and also in oscillating movement that results in modulation of the air stream. Such a source of acoustic energy essentially produces a spectral patterns such as illustrated in Figure 76.

APPENDIX J

THE IMPORTANCE OF THE VOCODER IN ACOUSTIC AND PHONETIC RESEARCH

One of the important tools developed and used in modern acoustics is the "vocoder", originally conceived at the beginning of World War II. The primary aims of its development were security of communication and reduction of bandwidth needed for transmission of speech.

To accomplish its effects the vocoder uses a bank of band-pass filters that can divide the speech wave into several bandwidths of frequency; each filter measures the energy concentration within its given range of frequency, and each filter emits a single wave which represents the composite energy of all sound waves of the original speech that fall within the band-pass filter range. Receiving equipment picks up the several waves representing energy and uses them to control the output of several oscillators, each assigned to a separate filter at the sending end. By combining the output of these oscillators into one wave and feeding this wave into a loudspeaker one can create a reasonable approxmation ot original speech. It should be noted that most vocoders and their adaptions measure speech frequency up to 4000 cycles per second although the actual speech wave has a range of 16,000 cycles per second or higher. The reason for this is that telephone netowrks had already demonstrated the possibility of transmitting recognizable gross characteristics of the human voice within a range not exceeding 4000 cycles.

Immediate uses of the vocoders were twofold. By "quantizing" energy of speech in accordance with the bandwidths of the filters, the vocoder reduced the amount of electronic information necessary for transmission of speech by a ratio of about 10:1. At the same time the use of filters and oscillators made it possible to "scramble" transmitted information by rapidly alternating the combi nations of carrier frequencies assigned to filters as well as to their respective receiving oscillators.

The first machine to utilize the process just described was later identified as the analogue channel vocoder to distinguish it from later adaptations discussed below. It differs from such adaptions in that it transmits information about the energy output of each filter continuously.

At present, there continues a discussion about the ideal bandwidth for vocoder filters, as well as the attenuation characteristics of the filters' "skirts." The problem is particularly important in the meaningful identification of information-bearing elements of the speech wave either in speech transmission or recognition, considered in the following

section. The greater the filter band-width, the less the resolution in the sound-wave energy; the smaller the filter band-width, however, the less possible it is to identify gross characteristics of the speech wave from energy in the harmonics of vocal flap frequency. The current trend is towards using filter bandwidths from 50 to 400 cycles per second and having "skirts" with a gradual (such as about 12db per octave) rather than a steep slope.

Such problems are the subject of a paper by G. G. M. Fant ("Acoustic analysis and synthesis of speech with applications to Swedish", Ericsson Technics 15, No. 1(1959)3-108), and of recent work at RCA (on contract with WADD). An alternative approach to the passing of speech through a bank of band-pass filters has been developed by the Federal Scientific (RADC contract No. AF 30 (602-1615). Since the subject of transformed waves is highly specialized, we refer the readers particularly interested in this subject to these papers.

An early modification of the analogue vocoder is the digitized vocoder. It is essentially the same as an analogue channel vocoder, but the energy output of the filter bank is sampled periodically and the level of this energy at sampling time is transmitted by pulses.

Later adaptions of the digital vocoder had two primary objectives. One goal was to transcribe more precise information about the information-bearing characteristics of speech; this need also contributed to the exact classification of phone groups and formants. It also produced the formant vocoders, which only use three or four filters to break up the entire speech wave. The second goal was to reduce significantly the rate of information transmission needed for speech communication; such needs led to the development of Caldwell Smith's modified vocoder, and it also gave an impetus to research into the exact acoustic classification of various phone groups. These efforts are discussed below.

Development of vocoding techniques gave scientists their first incentive to measure different spectral density distributions at different intervals of time. A particular impetus for this work was based on the differences in spectral density which seemed to be directly related to differences in sound that could be identified by ear. Such work led to the development of the spectrograph. A spectrograph is a special tool developed for a careful study of spectral density distributions of speech waves.

The spectrograph enabled scientists to produce graphic illustrations of formants as functions of the words articulated and of time. Formants are the main regions of peaks of spectral density envelopes. The voiced

portions of speech waves contain about three formants that are considered to convey significant information. These correspond to the three principal resonance chambers formed by various coupling of the throat and mouth cavities.

Conceptually, it should be noted that the precise characterization of formants is still a subject of discussion. In classic acoustic a formant is defined as the peak of the envelope of the spectral density distribution, but difficulty in determining these peaks precisely has led to inaccuracies of measurement. As an operational definition Peterson and Barney, whose work is discussed later in this section, have suggested applying the term formant to center of gravity of the spectral density in the regions of three decibels on either side of the density peak. This still leaves open the situation where the peaks are close together and the regions do not show a 3 db depression in the spectral density level. Graphically, the formant appears on a spectrogram as a dark band representing concentration of energy whose frequency level varies with time. This band may be divided into transitions (onglide and offglide) caused by movement of the articulatory organs from one sound to another and the production of a specific tone (steadystate). Specific investigation into the nature of formants has been particularly oriented toward identifying phones by the characteristic slope of their transitions and the level of their steady-states.

The first extensive published results of spectrographic analysis of formant energy distribution are reported in the textbook Visible Speech by Potter, Kopp, and Green. In this study the spectral densities of different speech wave-forms are displayed as functions of time. On the spectrograms each speech wave usually reveals from two to five distinct formants. Spectrographic transcriptions of speech were made and the transcriptions classified both according to the identities of the speakers and according to the sentences spoken for producing the wave forms.

Human observors, it was discovered, could actually read these patterns as whole sentences and even relate fragmented portions of the sound wave patterns to those portions of the spoken sentences that produced them. This was true even when the sentences were spoken by a variety of different speakers selected for the experiment.

The ability of speakers to recognize sounds by their gross energy distributions alone suggested the value of further investigation into the nature of such distribution, such as the classification of phones, discussed below. One of the objectives of such work was the reduction in rate of

information transmission for speech communication and the other was development of speech actuated machinery.

With digitized vocoders it is necessary to transmit information about energy levels at approximately twenty frequency ranges. If one could identify sounds through gross characteristics of their formants the number of ranges about which information need be transmitted might be reduced to three energy bands and the identification of pitch. This possibility was indicated by spectrographic studies of speech wherein formants were indicative of differences between the various sounds in speech.

Drs. Peterson and Barney of Bell Telephone Laboratories have also investigated the characteristics of steady state formants as identifiers of vowels of English.

The experimental procedure of this research was to analyze the spectral densities of specified English vowel sounds positioned between the consonant sounds "h" and "d." Experimenters used phonetic data from the enunciations of 75 select and trained speakers. Researchers took pains to obtain very careful enunciation; each speaker's vowel sounds were tested on a random audience before and after spectrographic recording to determine whether the sound enunciated was identifiable as a specific vowel.

Although the relevancy of experiments with careful articulation to the transcription of general rapid speech is still somewhat questionable, the Peterson, Barney data has indicated that levels of first and second formant frequencies gave reasonable indication of about 90% of the vowel sounds studied; the data also indicated formant overlap in regions representing two or more vowels. Such overlaps generally resulted from the different speech characteristics of different subjects. The experiments of Peterson and Barney thus indicate that it may be possible to represent steady-state vowel sounds solely by graphing the levels of formant frequency.

Studies of spectrograms such as those published in Visible Speech moreover indicate the possibility of representing consonant sounds by noise bursts preceding or following the vowel sounds, then adding a suitable transient to the formant levels between the consonant noise burst and the vowel steady state.

Moreover, the possibility of transmitting information about formants only was further substantiated by study of stylized patterns of speech spectrograms at Haskins Laboratories and synthesized speech produced by their Pattern Playback. [Researchers produced these

patterns by painting formant patterns on celluloid sheets graphed to measure frequency levels. When the selluloid passed across the Playback beams, the light reflected from the paint actuated the production of tones at various levels of frequency. Simulated sounds so produced were presented to human listeners for evaluation of the "quality" of the synthetic representation of phonemes studied for any test.)

By experimenting with the shape of formant patterns, the observing the resultant change in sound, Haskins Laboratories were able to generate a considerable amount of valuable information defining the exact phonetic changes produced by shifts in sound-wave energy. By changing the frequency and duration of onglides and offglides of vowel and consonant formants, for example, researchers found they could produce sounds similar to semi-vowels. Among other phonetic-acoustic characteristics about speech perception that were formalized by Haskins Laboratories are those of stop consonants and fricatives.

Such initial success in relating acoustic information about formants to phonetic perception led to two parallel efforts -- development of formant vocoders, and precise classification of phone groups.

Development of formant vocoders, using only three or four filters to track formants, was motivated by the desire to reduce the rate of information transmission for speech communication, as has been previously noted.

Refinements of formant vocoders have also contributed to the more efficient transmission of speech. Early formant vocoders operated by transmitting information about those filters which had spectral density maxima in their frequency range. The shifts in formant frequency with time were indicated only when such a change represented the shift of this maxima from the frequency range of one filter to that of an adjoining one. Hence the formant frequency changes were quantized according to the frequencies of the filters in the bank.

New heterodyne filters aim to control automatically the filter center frequency to correspond to the peak of the envelope of spectral density output; and also to incorporate automatic methods which would switch transmission from one filter to the next as necessitated by the movement in frequency of maxima of spectral density out put with time. While eliminating distortion introduced by quantization, such improvements also present a method for automatic measurement of formant frequency; information essential to automatic "phoneme" recognition using spectral density output. There is still room for improvement in the reliability with which formant tracking is accomplished by such methods.

Precise classification of phone groups, the other mahor field of investigation related to spectral analysis, provided even further opportunities for reducing the rate of transmitting information. Earlier work of linguists had already presented the possibility of specifying the speech of any Indo-European language by a small set of elemental sounds called phonemes. English is said to have between 35 to 42 phonemes. Assuming that phonemes are recognizable by certain gross characteristics, much as printed letters may be recognizable in human handwriting however distorted, it should be possible to build a machine to recognize these gross characteristics. Although an exaggeration, the metaphor provides a partial analogy to the initial thought processes that led to continuing research into the precise identification of phonetic sounds by their acoustic characteristics.

If such identification were feasible, the amount of information transmitted about speech would be reduced even more than through formant vocoders, since it would be possible to assign a code number to each phone and transmit only that code to actuate a receiver; in such a case the number of pulses or bits needed for transmission is said to be about twenty-four per second, compared with over 600 per second for digitized formant vocoders such as described in J. Flanagan's doctoral dissertation.

As yet, however, it has not been possible to discover a completely adequate method for grouping phones according to their acoustic data, despite the initial success of some limited purpose digit recognizers discussed in the following appendix. On reason for this is that the characteristics of each phoneme are modified by preceding and following phonemes, so that one phoneme group is likely to consist of several allophones, all sharing common characteristics of one sound, but each slightly different from its associates.

The analysis of speech into its possible allophones would in fact result in so large a number of possible sounds that the complete study of all their acoustic correlates would require considerably more time than that which has already been spent studying phone classes. considering the large number of allophones, it is indeed almost impossible to identify a separate phoneme by its acoustic patterns alone, particularly for the transmission of speech. Thus hampered, efforts at reducing speech bandwidth transmission through automatic phoneme recognition have not completely achieved their goal. Specific problems in phone grouping, more directly related to the practical development of speech recognizers, are discussed in the following appendix.

Faced with many unsolved problems in relation to the grouping of phones scientists have also worked on alternative methods of transmitting speech efficiently. One solution to the problem of phone grouping is to gather information about the speech wave at regular intervals of time without relation to phonemic segmentation. This is an extension of the technique of digitized vocoders previously mentioned; Caldwell Smith's modified vocoder presents a method of speech transmission that is considerably more efficient than most other methods that are ready for practical development.

In the vocoder developed by Smith, digital samples of quantized energy levels from each filter are fed to a temporary memory bank in a set frame instead of directly to a transmitter as in a digital channel vocoder. The frame in the temporary memory bank, representing the orderly output of each filter, is compared with other such frames stored in the permanent memory bank. The frame in the permanent memory bank that best matches the actual sample is transmitted to the receiver by its representative code.

At the receiver there is an identical set of stored frames representing possible digitized combinations of spectral energy; when it receives the coded signal, the receiver chooses the proper frame. The orderly stored representation of energies on the chosen frame then actuates the bank of oscillators; just as if the information about the total frame had been received as in a digital channel vocoder.

Although this process involves more steps than in the regular digitized vocoders, it eliminates the necessity of transmission of a large number of bits, since after assigning a code to the whole frame of digitized representations of energy levels, it is necessary to transmit only this code for the entire frame rather than all the energy levels of the various filters.

Speech processed by vocoders is often considered to lack "naturalness" of a speaker's "voice qualities." This is a result of synthesis of speech at the receiver, from information about energy outputs of about twenty filters that divide the speech spectrum, into a like number of frequency bands. Some effort has been made towards overcoming the above-mentioned limitation. In the beginning, information about the speaker's vocal flap frequency and its harmonics was presented along with the spectral density output of the band-pass filters. More recent devices, namely the voice excited vocoders developed at the Bell Telephone Laboratories, transmit low frequencies (i.e. those below one thousand cycles per second) directly (without vocoding) and the rest of the speech energy is vocoded for transmission purposes.

Such a system is reported to retain several of the "speaker's" voice qualities.

As is apparent from the review above the development of equipment for speech recognition has depended primarily on the efficient transmission of a recognizable sound-wave. Such investigation has produced a considerable amount of data useful in terms of a general purpose recognizer, at the same time certain aspects not directly relevant to efficient transmission of the speech wave have not been given much attention, and deserve further investigation for the uses of our model. There is, for instance, some question about the distortion caused by passing a speech wave through a bank of band-pass filters; for purposes of speech recognition our model may require additional data in the form of time-amplitude plots. The following Appendix considers some of the problems in the application of present equipment to the development of general speech recognizers.

APPENDIX K

THE DEVELOPMENT OF MACHINES FOR SPEECH RECOGNITION

Researchers investigating machinery for speech recognition relied heavily on techniques developed for vocoders, as has been mentioned; at the same time there was a primary difference in specific objectives between those wishing to reduce the amount of information needed for transmitting speech and those whose main interest lay in using available acoustic equipment to provide cues for speech transcription. Thus, while experiments in automatic transcription of speech are by no means independent of the methods and data employed in relation to vocoding technique, the goal of speech recognition should be kept separate both conceptually and procedurally.

The relationships and differences between the objectives of speech transmission and speech transcription help explain the state of our present data, for example. Construction of speech recognizers depends on the ability of electronic machines to "read" speech from the gross characteristics of sound-wave patterns. At present such patterns exist primarily in the form of spectrograms and time-amplitude plots. Spectrograms have a definite value in developing efficient speech band-width compression, since they form the major part of our acoustic information drawn from previous research.

There is a more direct coincidence of interests between speech transcription and transmission in the matter of grouping sounds for automatic recognition, although the exigencies of speech transcription may demand a more detailed analysis of suitable phone classification, segmentation, and normalization of duration than is presently provided by research related to vocoders. The probability of such conflicts should be kept in mind in the following discussion of actual experiments with speech recognizers. Initial efforts with speech recognizers depended on acoustic information available from spectrograms; sounds were related to the acoustic patterns they produced.

One of the first efforts to develop speech actuated machinery was the automatic digit recognizer of the Bell Telephone Laboratories, known as AUDREY. This device correlated the significant patterns of spoken digits, such as frequencies and durations of noise bursts and frequencies of formants. The success of such a recognizer was limited to about 65% accuracy; even when allowances were made for variation in the average formant frequencies of male and female speakers.

A more complex effort to transcribe speech was the automatic typewriter conceived by Dr. Olson of RCA Laboratories. For this development the information about location and duration of noise bursts,

the average levels of steady state formants as well as the average transition of formant frequencies between noise bursts and steady state formants were programmed for recognition of phonemes of a preselected vocabulary. The performance of this typewriter is much harder to evaluate than that of the digit recognizer discussed above; suffice it to say here that it was far from that needed for a phonetic typewriter.

Subsequent to the development of these recognizers Forgie and Forgie constructed an automatic vowel recognizer, that used characteristic patterns of the second formant as a cue to identification.

Various disadvantages of these machines suggested two needs - equipment for obtaining more precise information about the acoustic correlates of speech, and an orderly method of grouping phones for easier recognition.

At present the operation of automatic formant trackers is not sufficiently reliable to enable researchers to use for general recognition a considerable amount of published information about vowel sounds and formant transitions as a cue to consonants. The vowel recognizer of Forgie and Forgie, for instance, relied on a special definition of formants based on their levels of energy; such levels, while present in controlled experiments, might differ under different conditions of articulation by different speakers.

Additional work, moreover, is needed to specify the characteristics of unvoiced portions of speech that identify consonant sounds. Among recent investigators in this field are Haskins Laboratories, Fry and Denes at the University of London, and Docent Fant at the Royal Institute of Technology, Sweden. Their research has related the relative distribution of energy in preselected frequency bands to the consonant phonemes producing the patterns. Results also showed that formant energy distribution depended on the vowels that followed the consonants. Fry and Denes reported it was possible for a machine to recognize certain consonants by the above methods with an accuracy of 25% to 90%, but the majority of their score was closer to the lower limit.

Once researchers had achieved a limited success in identifying speech by the gross characteristics of the acoustic patterns, moreover, the value of more precise grouping of phones became apparent - for recognition as well as transmission of speech.

Acoustic recognition of vowels and acoustic recognition of consonants had been shown to depend on different criteria; past research with discrete speech had related vowels with steady state formant levels and consonants, with noise bursts, stop gaps, and transitions between

formants. Experiments with early recognizers, however, had also shown that phones could not be identified by their gross acoustic patterns alone. Exact identification seemed more likely to depend on an ordered analysis based on similarities and differences between various phone classes. Such an approach would provide one method for organizing undifferentiated masses of acoustic data into a form comprehensible for available electronic machinery. This approach would, moreover, provide a feasible method of correlating the phonetic and acoustic differences likely to occur between discrete and carefully articulated speech.

Grouping speech into categories based on their phonetic or phonemic characteristics has, accordingly, been the subject of much scientific interest. The particular value of such grouping is that it would substantially reduce the number of choices which a machine might need to make about sounds it perceived. Suppose that a transcribing machine hears the phone d in the word dic. It must distinguish between this word and other sound combinations which may be possible. There are approximately forty phone groups in the English language, and we assume that each phone modifies the pronunciation of adjacent phones. With this set of conditions the machine must theoretically choose between a number of sound combinations in the range of forty factorial. This is, of course, beyond the range of present computers.

Classification of sounds, however, reduced the process of choice to a series of separate decisions between whole categories of sound combinations; the machine decides whether a sound is voiced or unvoiced, whether it employs the nasal passages to accent resonance, whether it is articulated at the lips, alveolar ridge and teeth, or back of the mouth, and continues to make such decisions reducing the possible characteristics of the sound until it is possible to make one final decision identifying the phone. This process may reduce the number of possible decisions involved from a potential forty factorial to a number well within the capacities of modern computing machines. Although scientists have devoted considerable attention to such efficient classification, no general agreement exists about the number of acoustic characteristics necessary to identify a phone class for automatic recognition by machines.

A phonetic report on grouping titled "Preliminaries of Speech Analysis" was co-authored by Proffessors Halle, Jakobson and Fant. In this report they present twelve characteristics of phonemes, such as vocal chord resonance, nasal resonance, frictional noise and so on, which are either present or absent in any selected phoneme.

The first full scale effort for development of equipment for automatically grouping different sounds, however, seems to be that of

Professor Chang of North Eastern University. He tried to group sounds as vowels, semi-vowels, nasals, non-vocal plosives, vocal plosives, non-vocal fricatives with low or high energies, and vocal fricatives with low or high energies. This separation of groups was based primarily on spectral density distributions of phonemes in each group, on presence or absence of energy concentrations in the low frequency range, on over-all energy for any phoneme, and on existence or absence of a silence before the start of certain sounds.

Although several parts of Professor Chang's system were never built, his effort can be considered a partial success since he was able to separate certain groups of phonemes with over 90% reliability, whereas other groups could not exceed 70% reliability of separation.

Following reports of this work Dursch of IBM built a digit recognizer in which he grouped phonemes of spoken digits exxentially according to Chang's system. As a major innovation, however, Dursch used information in time amplitude waveforms of speech as well as studies of speech processed through band-pass filters, The performance of this recent digit recognizer is reported to be 95% to 97% reliable, comparing favarably with the 65% reliability reported for AUDREY. It is probable that further developments along this line of investigation will be made.

APPENDIX L

Relation of Available Information to Acoustic Correlates of Speech Necessary to the Development of our Model

1. Manner of Articulation

Data from Visible Speech, Truby, and Haskins Laboratories in the form of spectrograms confirm the following descriptions of manner of articulation:

a. Stops are denoted by stop gaps followed by energy bursts.

b. Spirants are characterized by continuous fricative energy.

c. Nasals are characterized by a special low-frequency nasal formant.

d. Sibilants are indicated by high frequency energy.

e. Affricates are denoted by a stop gap followed by high frequency energy.

The acoustic characteristics of laterals require the generation of additional acoustic data.

2. Place of Articulation

The primary acoustic evidence reflecting place of articulation would be the frequency levels of formant transition. Generation of further data about this dimension of our model requires that we obtain spectrograms of consonants such as [m, b, and p] with different manners of articulation but the same place, then measure the relative frequency of their onglides.

Data from Haskins with synthesized speech, it may be noted, indicates that the formant transitions following the consonant [m] start at the same frequency level as those for the consonant [b] although they have different manners of articulation. Such evidence, if also true of normal speech, would indicate that transitions are cues for place of articulation.

3. Voiced Nasal Resonances

Available information indicates that voiced nasal resonances such as m, n, and ŋ are characterized by a voice bar plus nasal formants.

4. Aspiration

The aspirated stops -- [p^c, t^c, k^c] -- are characterized acoustically by a stop burst followed by a period of friction.

5. Non-distinctive Consonant Differences Depending on Following Vowels

Evidence of different starting points for second formant onglides, particularly in the recent work of Bjorn Lindblom (Lindblom, 1963) show that the articulation of a consonant varies non-distinctively depending on the following phone.

6. Labialization of a Consonant Before a Rounded Vowel

Spectrograms of [s] before [w] show energy concentrated in the unvoiced portion at the level of [w]'s second formant, suggesting labialization of [s] before the labial [w].

7. Duration

We have measured the duration of vowels as suggested in the first section: such research suggests the importance of duration in identifying vowels, but further data is needed. We have no further data on duration differences caused by regional dialects, and this subject also requires investigation.

8. Intensity

Available information is in the form of broad-band spectrograms, which show only major variations in intensity. Further data may be available by measuring speech with time-amplitude plots.

9. Frequency

Broad-band spectrograms also do not indicate exact formant frequency; thus available data does not yield specific information about frequency. Necessary data requires the use of narrow-band spectrograms and cross sections, and of time-amplitude plots for further research.

10. Consonant Clusters

Truby (Truby, 1959) indicates that consonant clusters are co-articulated and that the articulation of the initial consonant may not affect the articulation of the following vowel. Further research in this area with time-amplitude plots and spectral analysis seems indicated.

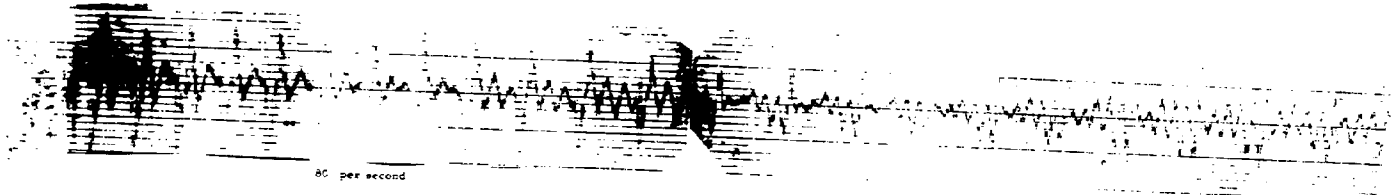
11. Nasalized Liquids and Semi-Vowels

Available data yields no information about the acoustic characteristics or existence of nasalized liquids and semi-vowels.

12. Classification of the y Phone.

It is worth noting, that although we mention on page 83 that "three years" and "three cars" might be treated as homonyms, analysis of time-amplitude plots for Speaker 5 (of our own data) (See Figures 73 and 74) indicate that y can be detected. In "three years"

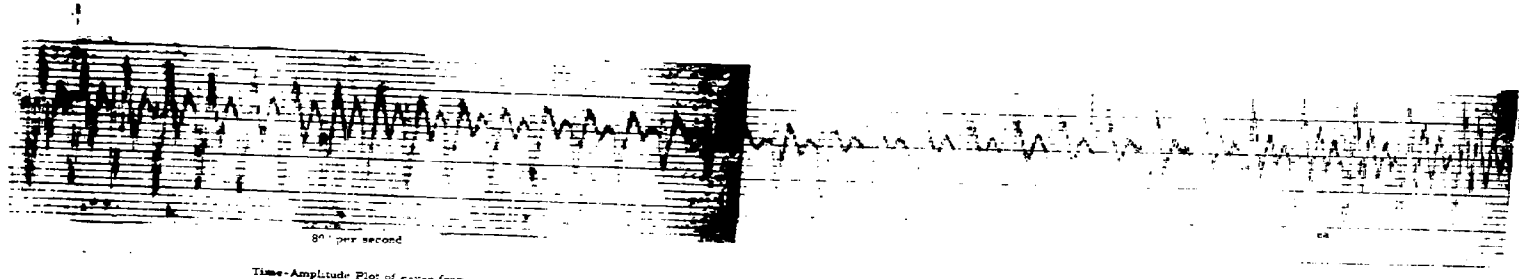
(Figure 74) there is no change in the frequency pattern, but a y is indicated by a decrease in the intensity of a portion of the waveform. "Three ears" (Figure 73) gives no such indication; there is, in other words, no detectible y.



80 per second

Time-Amplitude Plot of one from
"three years" by Speaker 5

Figure 73



80 per second

Time-Amplitude Plot of one from
"three years" by Speaker 5

Figure 74

APPENDIX M

MEASUREMENTS MADE FOR DETERMINING ACOUSTIC CHARACTERISTICS OF SPEECH

We performed measurements on spectrograms to determine the frequency of occurrence of various parts of the speech wave such as the second formant onglide during the articulation of consonant-vowel combinations. The extreme variations in the spectrum caused by the articulation of one consonant with different vowels suggests the acoustic interdependence of consonant and vowel articulation as indicated by their organization in our model through placement on different planes.

Data was taken exclusively from Visible Speech; to obtain it the following method was used: measurements of illustrations of spectrograms were made with a ruler whose smallest division was 1/16." These measurements were then tabulated with a scale given on page 12 of Visible Speech. Height of the illustrations was approximately 15/16 of an inch and by the given scale 1/16" represented 233 cycles or 23 milliseconds. Thus our measurements of frequency and duration are representative rather than definitive, as there is considerable margin for error. In certain cases, it was impossible to measure every aspect of the onglide, steady-state and offglide.

We also measured some data published by Truby indicating the effect of final consonants on vowel frequency and duration; although minimal pairs would have been desirable for compilation of this data, they were not available. In these measurements we used a ruler whose smallest unit was one-eightieth of an inch. Illustrations from Truby, however, are smaller than those from Visible Speech; according to his scale one-eightieth of an inch would equal 103 cycles and one sixteenth of an inch would equal 515 cycles.

APPENDIX N

Finally, we performed measurements of the proportionate relationships between onglide, steady-state, and off-glide for four words taken from the work of Lehiste and Peterson. The investigators include no scale for their illustration, but the relative durations of [ɪ] and [i] form one important criterion for distinguishing between these sounds; thus the data from Lehiste and Peterson shows that duration is an element important to the evaluation of acoustic data by a general speech transcriber.

APPENDIX C

Rules of Euphonic Combination Supportec by Acoustic Evidence in the Data Studied

Listed below are the rules of euphonic combination for which we have been able to find available acoustic evidence in the form of spectrograms. Numbers after the quoted sentences, unless otherwise noted, refer to the page number of photographic illustrations in Visible Speech.

RULE	ILLUSTRATIVE SPECTROGRAM	SOUND SUBSTITUTION	ACOUSTIC EVIDENCE
1. A voiceless consonant becomes voiced next to a voiced consonant, when a final voiceless consonant precedes a word beginning with a voiced consonant. (Appendix H. III)	"Show us that beige shirt." (149)	The t of that is voiced.	There is a voice bar in the stop gap of t.
2. A voiceless consonant becomes voiced between vowels, when it occurs at the end of a word and the next word begins with a vowel. (Appendix H. III.)	"She made me my hat." (173) "He said something about it." (188)	The h of hat is voiced. The t of about is voiced.	There is a voice bar in this consonant. There is a voice bar in the stop gap of t.
3. A voiced consonant can become voiceless if it occurs before a voiceless stop, spirant, or sibilant (Appendix H. III)	"Has she found a sound?" (201) "This is such a big church." (p. 156)	The final consonant of has, which normally is [z], is voiceless here. The "s" of is, which is normally [z], is voiceless here.	There is no voice bar in this consonant. There is no voice bar for the "s" of is.

RULE	ILLUSTRATIVE SPECTROGRAM	SOUND SUBSTITUTION	ACOUSTIC
4. t becomes p everywhere except after a spirant or sibilant (Appendix H. II)	"Did you talk about it?" (99)	The final consonant of about is a stop and not a stop.	There is no st
5. All voiced unaspirated stops become spirants (Appendix H. II)	"Don't stand behind him." (200)	The d of stand is a spirant; the spectrogram does not precisely the sound is	There is no st
6. A voiceless alveolar stop plus y becomes [č] in English (Appendix H.III)	"but you" (255)	There is a between the vowel and you.	This spectrog patterns simil of the
7. Before or after a dental, an alveolar may become a dental, as in health and width. (Appendix H.VII)	"Can the man be certain then." (182)	The fina has a denta place of articulat	There is no s second formant the same place of articulation.
8. A final consonant of one word may be attached to an initial vowel of the following word. (Appendix H. IX)	"Could you take it back." (100) "Top it." (92)	The k of take is attached to the and it. The p of is attached to the vowel of it.	There is no p between the e of the followi There is no p between the e of the followi
9. When a nasal is followed by a voiceless sibilant or spirant, a voiceless stop may be inserted between them. (Appendix H. IX)	"once" (195) "length" (195)	There is a between the n and the s. There is a between the n and the p.	There is a st There is a st

RULE	ILLUSTRATIVE SPECTROGRAM	SOUND SUBSTITUTION	ACOUSTIC EVIDENCE
10. A vowel may be replaced by a visarga vowel. (Appendix H. IX)	"Fine weather is infrequent in December." (Denes, p. 27).	The final vowel of "December" (which in this speaker's dialect is not followed by r) is a visarga vowel.	The [ə] in "December" has weaker formants with broader bandwidths than does the [ə] in weather, where the r is pronounced.
11. Final d may drop out after n. (Appendix H. IX)			There are no examples in Visible Speech in which this actually occurs, but almost every d after n is much shorter than any d which is not part of a consonant cluster. The final d of thousand (p. 199) has a stop-gap approximately one-fourth as long as the stop-gap of the final d in bad (p. 90). It is quite probable that such a short consonant should drop completely.
12. A glottal stop may occur before any initial vowel. (Appendix H. IX)	"Save us a set of those." (142) "Is it among these?" (188) "Begins at ten." (181)	Glottal stop before "of". Glottal stop before "among." (Maybe before "is" and "it" also.) Glottal stop before "at".	The initial vowel in of begins with an energy burst. The initial vowel of among begins with an energy burst. The initial vowel of at begins with an energy burst. There are other sentences in Visible Speech which appear to have glottal stops in them, but they are not all as clear as the ones cited above. Since a glottal stop has no formant transitions intrinsically associated with it (see (Visible Speech, p. 80), it appears on a spectrogram simply as a stop gap without a voice bar followed by a stop burst.

RULE	ILLUSTRATIVE SPECTROGRAM	SOUND SUBSTITUTION	ACOUSTIC EVIDENCE
12 (Cont'd.)			If a pause precedes the glottal stop, the stop gap cannot be distinguished, and if the stop burst is weak, it cannot be identified. For these reasons a glottal stop does not always show up clearly on a spectrogram. Thus in our second example above, there may be other glottal stops in the sentence, but we cannot tell from looking at the spectrogram of Visible Speech.
13. A voiced consonant in a voiced environment may become voiceless and sometimes aspirated Appendix H. IX	"She put rouge on her face." (149) "We can begin when you "We can begin when you come." (250)	The of rouge is voiceless throughout more than half of its duration. The g of begin is voiceless throughout part of the stop closure. It is also aspirated.	There is no voice bar in part of the of rouge. The g of begin has no voice bar and the stop burst is followed by aspiration.

APPENDIX P

DISCUSSION OF THE RULES OF EUPHONIC COMBINATION SUPPORTED BY PHONETIC TRANSCRIPTION

A. Texts Used

The passages used in our analysis of continuous speech are taken from two sources as mentioned in the main body of our text: a record by Jackie Gleason (Decca 27684) and tapes of natural conversation. One side of the Jackie Gleason record is entitled, "What Is a Boy?" and the other side, "What Is a Girl." The tape recordings of conversation were made by Dr. J. M. Pickett of the Air Force Cambridge Research Laboratories; Dr. Pickett kindly let us make copies of them.

Each side of the Jackie Gleason record has been divided into four parts and a number assigned to each part. Every time we cite a passage from this record, we give the number of the part in which it appears.

On the side called, "What Is a Boy" Part I begins with "Between" and ends with "and Heaven." Part II begins with "protects them" and ends with "Paul Bunyan, the". Part III begins with "shyness of" and ends with "nobody else can". Part IV begins with "cram into" and ends with "Dad."

On the side called, "What Is a Girl", Part I begins with "little" and ends with "special look". Part II begins with "in her eyes" and ends with "softness of a". Part III begins with "kitten" and ends with

"flirtatious". Part IV begins with "when she" and ends with "of all".

The passages from the AFCRL tapes which contain our examples are quoted below; before each passage, we give a brief description of its context. Each passage has been assigned a number.

Passage I

Conversation about anechoic chamber with girl who said she was majoring in the psychology of education.

Female voice: 's fascinating. It looks like an attic.

Male voice: Some people come in, first remark they make is, "My ears seem to feel funny.

Female voice: My ears didn't feel funny but um uh speech sounds a little bit different, sort of muffled.

Male voice: Yes, if you uh clap (clapping sound).

Female voice: Yeah.

Male voice: Sort of . . .

Female voice: Yeah, 'ts funny.

Passage II

Conversation about the word list with the girl who said she was majoring in government.

Male voice: We use those in a way to calibrate our speech system, since we right now can't put a little thi . . . something like a voltmeter on, we . . . we have to test our system with speech itself.

Passage III

Conversation about regional accents with girl majoring in government.

Male voice: Well, don't you sometimes stick r's in when...

Female voice: Once in a while.

Passage IV

Conversation about courses required for major in government.

Male voice: Is this partly city planning? I. . . I don't really know...

Female voice: No.

Male voice: Theory of government?

Female voice: Uh, well, this year it's kind of a general. . .

B. Rules of Euphonic Combination Substantiated or Suggested by Phonetic Transcription

Numbers in parentheses after rules cited in this division refer to the appendix where the rule may be found. In examples cited below these rules, the bracketed letters represent phonetic symbols according to the modified standard International Phonetics Association Alphabet (In accordance with the consonant categories listed in Appendix B). The phonetic statement is followed by an example drawn from our transcription. The notations "Boy I" "Girl II" , or "AFCRL tapes III" refer respectively to specific texts of "What is a Boy?", "What is a Girl?" or the recordings

of conversation from Hanscom Air Force Base. Thus the notation, [n̪] becomes [n] between the (Boy I), means that an alveolar n becomes a palatal n as exemplified in our transcription of the phrase between the contained in the first textual segment analyzed from "What is a Boy?"

1. Change in Place of Articulation:

(a) Before or after a dental an alveolar may become a dental.

[n̪] becomes [n] between the (Boy I)

[n̪] becomes [n] with noise (Boy I)

(b) Before or after a palatal consonant, a dental or alveolar nasal may become a palatal (Appendix H. I).

[n̪] becomes [ñ] enjoy (Boy I)

[n̪] becomes [ñ] when you come home (Boy I). It is interesting to note that the phrase when you are busy has [n̪] rather than [ñ], although the environment is the same.

2. Change in Resonances:

(a) A voiceless consonant may become voiced next to a voiced consonant or between vowels (Appendix H. III)

[k] becomes [g] comic books (Boy III) The final consonant of comic is [g] rather than [k].

[s] becomes [z] across the (Boy III)

[t] becomes [d] top it all (Girl III) The final consonant of it is [d].

[t] becomes [d] fascinating (AFCRL tapes I)

[t] becomes [d] attic (AFCRL tapes I)

[t] becomes [d] little (AFCRL tapes I)

[s] becomes [z] once (AFCRL tapes III)

(b) A voiced consonant may become voiceless next to a voiceless consonant (Appendix H. III).

[b] becomes [p] absolutely (Girl IV)

[v] becomes [f] of string (Boy III)

3. Sound Drop-outs

(a) An alveolar stop between two consonants may drop out:

after [n] (See Appendix H. IV).

and the [æn ði] (Boy I)

and colors [æn kʌləz] (Boy I)

didn't feel [dɪdn fil] (AFCRL tapes I)

sounds a [saʊn zə] (AFCRL tapes I)

after [s]

best clothes [bɛs klovz] (Girl I)

must not [mʌs nat] (Girl IV)

first grade [fɪrs greɪd] (Girl III)

after another stop

protects [pro tɛks] (Girl II)

after a spirant

softness [sɔf nɪs] (Girl II)

(b) An alveolar stop before an affricate may drop out. (Appendix H. IV)

before [č]

straight chairs [streɪ čɛrz] (Girl III)

cat chasing [kæ ˇčeɪ sɪŋ] (Boy IV)

(c) Initial h may drop in an unstressed syllable Appendix H. II)

after [n]

in her [ɜ nər] (Girl II)

after [s]

grasshopper [græs sa pə] (Girl II)

after [k]

lock him [la kɪm] (Boy IV)

4. Shortening of long consonants

(a) When two identical consonants come together they form one long consonant which may be shortened to the normal length of a singal consonant. This rule also affects long consonants produced by other rules of euphonic combination. (Appendix U. IX)

special look [spɛʃə lʊk] (Girl I)

makes something [meɪk sʌm ˆθɪŋ] (Boy ...)

gets so [gɛt so] (Boy III)

gumdrops six cents [gʌm drap sɪk sɛnts] (Boy IV)

supersonic code [su pə sa ni kovd] (Boy IV)

(b) Combinations of identical consonants which are the result of other rules:

next door [nɛk. stɔr]; [d] becomes [t] and the long [t] is shortened (Girl III)

knives saws [naɪv sɔz]; [z] becomes [s] and the long [s] is shortened. (Boy III)

trains Saturday	[trɛɪn sæ tə dɛɪ]	[z] becomes [s] and the long [s] is shortened. (Boy III)
is so	[ɪ so]	[z] becomes [s] and the long [s] is shortened. (Boy III)
bedtime	[bɛ taɪm]	[d] becomes [t] and the long [t] is shortened. (Boy III)
ears seem	[ɪr sim]	[z] becomes [s] and the long [s] is shortened. (AFCRL tapes I)
bit different	[bɪ dɪ frʌnt]	[t] becomes [d] and the long [d] is shortened. (AFCRL tapes I).

5. Euphonic Combination of Final Consonants with Following Phones:

The final consonant of one word may attach itself to the initial phone of the following word. (Appendix V. IX)

We have so many examples for this type of sound combination that we have not listed them all. Examples of such combination from careful speech cited below are found in part (Boy I) of the Jackie Gleason record. For the AFCRL tapes of conversational speech we list all examples found in passage I.

Jackie Gleason record:

innocence of	[ɪnosɪn səv]	babyhood and	[bɛɪbihʊ dæn]
find a	[faɪn də]	come in assorted	[kʌ mɪ nəsɔətɪd]
weights and	[weɪt sən]	second of every	[sɛkən də vɛvri]
of every	[ə vɛvri]	hour of every	[aʊ ɪə vɛvri]
their only	[thɛ ɪɔnly]	them off	[thɛ mɪf]
boys are	[boɪ zɛ]	found everywhere	[faʊn dɛvriwɛɪ]

top of underneath inside of	[ta pʌ vʌndəni ɒɪnsaɪ]	climbing on	[klaɪmɪ ŋan]
running around or	[ɹʌnɪ ŋaraun dɔr]	mothers love	[mʌðəz zlʌv]
sisters and	[sɪstə zən]	adults ignore	[ʔæ dʌl tsɪgnɔə]

AFCRL tapes:

's fascinating	[sfæ si nɛr dɪŋ]	looks like an attic	[lʊk slaɪ kə næ dɪk]
come in	[kə mɪn]	first remark	[fərs tri mark]
make is	[mɛɪ kɪz]	ears didn't	[iə zdɪ dn]
sounds a	[saʊn zə]	if you	[ɪ fyu]
sort of	[sɔr tə]	's funny	[tsfʌ ni]

6. Glottal stops

(a) A glottal stop may be introduced before a word starting with a vowel (Appendix II.IX)

[æ] becomes [ʔæ] adults (Boy I) There is a glottal stop before the initial vowel of adults.

(b) A glottal stop may be substituted for [t] before a labial consonant (Appendix II.X)

[tm] becomes [ʔm] voltmeter (AFCRL tapes II) The consonant before the m is a glottal stop rather than a [t].

7. Consonant clusters involving ɹ: The combination of [s] followed by [ɹ] may become [ʃ]. (Appendix II.II)

this year [ðɪ ʃir] (AFCRL tapes IV)

8. **Treatment of [r] in New England Dialects**

(a) **An [r] after a vowel may drop except when [r] stands at the end of a word and the next word begins with a vowel.** (Appendix H, IX)

(1) Loss of [r] between a vowel and a consonant:

their last	[ðɛ lɛ̃st]	(Boy I)
fire cracker	[faɪə krækə]	(Boy III)

(2) Retention of [r] between two vowels:

their only	[ðɛɹ roʊn li]	(Boy I)
fire engines	[faɹ rɛn ʤənz]	(Boy III)

(b) In some New England dialects [r] may be inserted between a word ending with a vowel and a word beginning with a vowel. (Appendix H, IX)

law of	[lɔ ɹʌv]	(Girl II)

APPENDIX Q

Discussion of w phone class:

Researchers at Haskins Laboratories have reported that an initial onglide of at least 50 milliseconds duration that begins at or close to the [w] locus will be perceived as a [w] (Liberman, Delattre, Gerstman, and Cooper, 1956). Observations like this are important to the successful construction of stylized formant patterns. But we object to the assumption that such stylized patterns, produced through mechanical methods, can be used to provide information about the acoustic characteristics of actual speech. After a review of the data we recently generated, however, we are able to make a conclusion about the function of onglide duration in the determination of w: for, the duration of the onglide was not found to be consistently greater for the phrase with w than for the phrase without it. For Speakers 1 and 5 "no ax" has the shorter onglide, if we do not include the duration of the pause, but for Speaker 2 "no wax" is the shorter onglide (See Figures 21 - 26). So the initial onglide of Speakers 1 and 5 tend to validate the Pattern Playback constructions of Haskins Laboratories: but the example of Speaker 2 makes it very difficult for us to consider initial onglide duration measurement solution for the identification of w -- as a rule.

Haskins tends to define the characteristics of speech by evaluating the judgement of listeners listening to the speech patterns Haskins produces on the Pattern Playback. But the perception of speech patterns does not necessarily provide us with information about the realities of speech production: our onglide duration analysis of the w phone indicates this.

The attempt to comprehend the acoustic correlate(s) of w is further complicated by the fact that we have recorded an example of initial [w] without any apparent steady-state: this is in the word "will" uttered by Speaker 5 in the sentence "Will you help us?" (See Figure 41). It appears that [w] is marked by a long steady state in some environments and a long onglide in others. Further research is needed to determine in which cases it has a long steady state and in which a long onglide. Before our work with transitional information pertinent to [w] is complete we hope to conclusively, Conjecturally, it is possible that all initial [w]'s are marked by long onglides while all non-initial [w]'s are marked by long steady-states.

ACKNOWLEDGEMENTS

We gratefully acknowledge the interest of Mr. Caldwell Smith and his recognition of the importance of the present research; thanks are also due to the Air Force Cambridge Research Laboratories for initiating this project.

We further acknowledge the help, information and guidance given by Mr. Smith, through discussion of other work on speech processing and its relation to our model concepts. He also made it possible for us to process our speech data at AFCRL.

We also acknowledge the assistance of Mr. Weiant Wathen-Dunn, Dr. J.M. Pickett, and Mr. Philip Liberman of the AFCRL speech Research Laboratory, who provided us with tapes of normal conversational speech and helpful comments pertaining to the description of normal conversational speech, as distinguished from the enunciation of isolated words.

The time-amplitude plots of speech were made at the Minneapolis Honeywell facilities in Boston by arrangement with Mr. McCarty and Mr. Mara. Their cooperation and help is gratefully recognized.

We also appreciate the cooperation of the Harvard Dramatic Society, who provided speakers for the generation of our data.

PERSONNEL

The principal contributor to this effort was B. V. Bhimani; the linguistic aspects of speech were studied by Mrs. M. F. D. Degges. The mathematical formulations and symbolic representations were performed and checked by Mr. Frank Rubin and Mr. Mark Douma. The writing of the monthly reports was done by Mr. Gordon Milde, Mr. Gerald Hillman, and Miss Margery Lowenberg. Mrs. Joeth Barker also aided in the preparation of monthly reports; in addition, she organized and edited the final report.

BIBLIOGRAPHY

Bhandarkar, R. G. First Book of Sanskrit, (33rd ed, Bombay, 1957.)

Bloch, Bernard, "Phonemic Overlapping," American Speech, (1940), 16

Bloomfield, Leonard, Language, New York (1933)

Bolinger, Dwight, "A Theory of Pitch Accent in English" Word. (1958), 14

Buck, Carl D., Comparative Grammar of Greek and Latin, Chicago (1933)

Classe, Andre, Rhythm of English Prose, Oxford (1939).

Cooper, F. S.: Delattre, P. C.: Liberman, A. M. : Borst, J. M.: and Gerstman, L. J., "Some Experiments in the Perception of Synthetic Speech Sounds." Journal of the Acoustical Society of America (1952), 24

Delattre, P. C.: Liberman, A. M.: anc Cooper, F. S., "Acoustic Loci and Transitional Cues for Consonants," Journal of the Acoustical Society of America, 27:4 (1955).

Denes, P., "Effect of Duration on the Perception of Voicing," Journal of the Acoustical Society of America, (1955), 27

von Essen, Otto, Allgemeine und Angewandte Phonetik, Berline (1953).

Fant, Gunnar, Acoustic Theory of Speech Production, 15 - Gravenhage, Mouton & Co., (1960).

Fant, Gunnar, and Lindblom, Bjorn, "Studies of Minimal Speech Sound Units." Speech Transmission Laboratory, Quarterly Progress and Status Report, 2/1961, Royal Institute of Technology, Stockholm, pp. 1 - 11.

"First Grammatical Treatise," edited and translated by Elinar Haugen, Language, Supplement to Vol. 26: 4, Oct. - Dec., 1950.

Fischer-Jørgenson, Eli, "Acoustic Analysis of Stop Consonants," Miscellanea Phonetica, (1954), 2.

Fourquet, J., Les mutations consonantiques du germanique, Paris (1948).

Fry, Dennis, "Experiments in the Perception of Stress," Language and Speech, No. 2, (1958,)1.

Harrell, Richard, "Some English Nasal Articulations," Language (1958), 34

Harris, Katherine S., "Cues for the Discrimination of American English Fricatives in Spoken Syllables," Language and Speech, (1958), 1

Heusler, Andreas, Altislandisches Elementarbuch, 4th edition, Heidelberg (1950), (reprint of 3rd edition, 1931).

Jackson, Kenneth, Language and History in Early Britain, Edinburgh (1953).

Lehiste, Ilse, An Acoustic-Phonetic Study of Internal Open Juncture, supplement to Phonetica, (1960), 5.

Lehiste, Ilse, and Peterson, Gordon E., "Duration of Syllable Nuclei in English," Studies in Syllable Nuclei 2, Speech Research Laboratory, Ann Arbor, Michigan, (1960).

Lehiste, Ilse, and Peterson, Gordon E., "Transitions, Glides, and Diphthongs," Studies in Syllable Nuclei 2, Speech Research Laboratory, Ann Arbor, Michigan, (1960).

Lewis, Henry, and Pedersen, Holger, A Concise Comparative Celtic Grammar, Gottingen, (1937).

Liberman, A. M.; Delattre, P.; and Cooper, F. S., "The Role of Selected Stimulus-Variables in the Perception of the Unvoiced Stop Consonants," American Journal of [illegible] (1952), 55.

Liberman, A. M.; Delattre, P.; and Cooper, F. S., "Some Cues for the Distinction between Voiced and Voiceless Stops in Initial Position," Language and Speech, (1958), 1.

Liberman, A. M.; Delattre, P.; Gerstman, L. S.; and Cooper, F. S., "Tempo of Frequency Change as a Cue for Distinguishing Classes of Speech Sounds," Journal of Experimental Psychology, (1956), 52.

Lindblom, Bjorn, On Vowel Reduction, Report No. 29, Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, (1963).

Lisker, Leigh, "Closure Duration and the Intervocalic Voiced-Voiceless Distinction in English," Language, (1957) 33.

Martinet, Andre, "Function, Structure, and Sound Change," Word, Vol. 8, 1952, 1 - 32.

Martinet, Andre, Economie des changements phonetiques, Berne (1955)

Menzerath, P., and de Lacerda, A., Koartikulation, Steurung, und Lautabgrenzung, Berlin and Bonn (1933).

Meyer, Ernst A., Englische Lautdaur, Uppsala, (1903)

Niedermann, Max, Precis de phonetique historique du latin, 3rd edition, Paris (1953).

O'Connor, J. D.: Gerstman, L. J.: Liberman, A. M. Delattre, P. C.: and Cooper, F. S., "Acoustic Cues for the Perception of Initial /w, r, j, l/ in English," Word, (1957), 13.

Peterson, Gordon E. and Barney, Harold L., "Control Methods Used in a Study of the Vowels," Journal of the Acoustical Society of America, Vol. 24, 1952, 175 - 184.

Sharf, Donald J., "Duration of Post-stress Intervocalic Stops and Preceding Vowels," Language and Speech, (1962), 5

Streitberg, W., Urgermanische Grammatik, Heidelberg (1895)

Thomas, Charles K., An Introduction to the Phonetics of American English, New York (1947).

Thurneysen, Rudolf, translated by Binchy, D. A. and Bergin, Osborn A Grammar of Old Irish, Dublin (1947)

Whitney, William D., Sanskrit Grammar, Cambridge, (1889).

LIST A

Code	Organization	No. of Copies
AF 5	AFMTC (AFMTC Tech Library-MU135 Patrick AFB, Fla. -for unclassified material	1
	AFMTC (MTBAT) Patrick AFB, Fla. -for classified material	
AF 18	AUL Maxwell AFB, Ala.	1
AF 32	OAR (RROS, Col. John R. Fowler) Tempo D 4th and Indepindence Avenue Wash 25, D. C.	1
AF 33	AFOSR, OAR (SRYP) Tempo D 4th and Independence Avenue Wash 25, D. C.	1
AF 43	ASD (ASNXRR) Wright-Patterson AFB, Ohio	1
AF 124	RADC (RAALD) Attn: Documents Library Griffiss AFB, New York	1
AF 139	AF Missile Development Center (MDGRT) Holloman AFB, New Mexico	1
AF 314	Hq. OAR (RRY) Attn: James A. Fava, Col. USAF Wash 25, D. C.	1
Ar 5	Commanding General USASRDL Ft. Monmouth, New Jersey Attn: Tech. Doc. Ctr. SIGR /SL-ADT	1
Ar 9	Department of the Army Office of the Chief Signal Officer Wash 25, D. C. Attn: SEGRD-4a-2	1
Ar 50	Commanding Officer Attn: ORDTL -012 Diamond Ordnance Fuze Laboratories Wash 25, D. C.	1

Code	Organization	No. of Copies
Ar 67	Redstone Scientific Information Center U. S. Army Missile Command Redstone Arsenal, Alabama	1
G 2	Defense Documentation Center (DDC) Cameron Station Alexandria, Virginia	20
G 31	Office of Scientific Intelligence Central Intelligence Agency 2430 E Street, N.W. Wash 25, D. C.	1
G 68	Scientific and Technical Information Facility Attn: NASA Representative (S-AK-DL) P. O. Box 5700 Bethesda, Maryland	1
G 109	Director Langley Research Center National Aeronautics and Space Administration Langley Field, Virginia	1
M 6	AFCRL, OAR (CRXRA - Stop 39) L. G. Hanscom Field Bedford, Mass.	20
M 77	Hq. AFCRL, OAR (CRTE, M. B. Gilbert) L. G. Hanscom Field, Bedford, Mass.	1
M 83	Hq. AFCRL, OAR (CRTPM) L. G. Hanscom Field, Bedford, Mass.	1
N 9	Chief, Bureau of Naval Weapons Department of the Navy Washington 25, D. C. Attn: DLI-31	2
N 29	Director (Code 2027) U. S. Naval Research Laboratory Wash 25, D. C.	2
R 292	Director, USAF Project RAND The Rand Corporation 1700 Main Street Santa Monica, California THRU: AF Liaison Office	1

Code	Organization	No. of Copies
U 443	Institute of Science and Technology The University of Michigan Post Office Box 618 Ann Arbor, Michgan Attn: BAMIRAC Library	1
AF 318	Aero Res. Lab. (OAR) AROL Lib. AFL 2292, Bldg. 450 Wright-Patterson AFB, Ohio	1
Ar 107	U. S. Army Aviation Human Research Unit U. S. Continental Army Command P. O. Box 428, Fort Rucker, Alabama Attn: Maj. Arne H. Eliasson	1
G 8	Library Boulder Laboratories National Bureau of Standards Boulder, Colorado	2
M 63	Institute of the Aerospace Sciences, Inc. 2 East 64th Street New York 21, New York Attn: Librarian	1
M 84	AFCRL, OAR (CRXR, J. R. Marple) L. G. Hanscom Field, Bedford, Mass	1
N 73	Office of Naval Research Branch Office, London Navy 100, Box 39 F. P. O. New York, N. Y.	5
U 32	Massachusetts Institute of Technology Research Laboratory Building 26, Room 327 Cambridge 39, Mass. Attn: John H. Hewitt	1
U 431	Alderman Library University of Virginia Charlottesville, Virginia.	1

Code	Organization	No. of Copies
G 6	Scientific Information Officer British Defence Staffs Defence Research Staff British Embassy 3100 Massachusetts Avenue, N. W. Washington 8, D. C.	3
G 9	Defence Research Member Canadian Joint Staff 2450 Massachusetts Avenue, N. W. Washington 8, D. C.	3

List G

U 249	Professor Roman Jakobson Massachusetts Institute of Technology 377 Massachusetts Avenue Cambridge 39, Mass.	1
U 250	Massachusetts Institute of Technology 377 Massachusetts Avenue Cambridge 39, Mass. Attn: Dr. Kenneth N. Stevens Research Laboratory of Electronics	1
U 251	Joint Speech Research Unit Eastcote Road Ruislip, Middlesex, England Attn: Dr. J. Swaffield	1
	AFCRL, OAR (CRBS, Caldwell P. Smith) L. G. Hanscom Field, Bedford, Mass.	64

LIST G

Code	Organization	No. of Copies
AF 1	Hq. ESD (AFSC) Operational Applications Laboratory Attn: (ESRH) L. G. Hanscom Field, Bedford, Mass.	1
AF 20	AFSC (SCSED) Andrews AFB Wash. 25, D. C.	1
AF 21	Hq. ESD (AFSC) Attn: ESRD, Lt. Col. Sidney W. Sheets L. G. Hanscom Field, Bedford, Mass	1
AF 22	AFORQ - OS/C Attn: Major E. T. Garrett Room 5C1067, The Pentagon Wash. 25, D. C.	1
AF 37	USAF Security Service (SED - 2) San Antonio, Texas	1
AF 51	Hq. USAF AFOCC - BB (Attn: Major W. K. Winbigler - Room 5B486) Pentagon, Wash. 25, D. C.	1
AF 74	RADC (RCUED, Mr. Richard C. Benoit, Jr.) Griffiss AFB, N. Y.	1
AF 123	AFAL (AVWC) Wright - Patterson AFB, Ohio 45433	1
AF 320	A. F. Electronic Systems Division (ESSDES, Lt. Col. T. Warns) 124 Trapelo Road Waltham 54 Mass	1
AF 321	RADC (RCDAI) Griffiss AFB, N. Y.	1
Ar 6	Office of the Chief Signal Officer Command and Control Systems Division SIGSD - 14, Attn: Mr. A. L. Ware Wash 25, D. C.	1
Ar 20	Chief, U. S. Army Security Agency Arlington Hall Station Arlington 12, Virginia Attn: ACofS, G4, TI, Section	1
AF 52	ESD (ESRC, Lt. Elrod, Stop 35) L. G. Hanscom Field, Bedford, Mass.	1

Code	Organization	No. of Copies
Ar 21	Commanding General USASRDL Fort Monmouth, New Jersey Attn: SIGFM/EL-NX-4, Mr. R. E. Lacy	1
Ar 85	Commanding General USASRDL Fort Monmouth, New Jersey Attn: SIGFM/EL-NRM	1
G 16	Defense Communications Agency Attn: Code 433 Wash 25, D. C.	1
G 30	Director National Security Agency Fort George G. Meade, Maryland Attn: R12, Mr. Rosenbloom	1
G 32	GPO Research Station Dollis Hill London NW2, England	1
G 127	Defense Communications Agency (321) Attn: M/Sgt. David M. Humphrey Wash 25, D. C.	1
G 128	Defense Communications Agency Code 722, Attn: Robert M. Scott Wash 25, D. C.	1
I 9	Bell Telephone Laboratories, Inc. Murray Hill, New Jersey Attn: Dr. Manfred Schroeder	1
I 13	Bell Telephone Laboratories, Inc. Whippany Laboratory Whippany, New Jersey Attn: Technical Information Library	1
I 52	Haskins Laboratories, Inc. 305 East 43rd Street New York 17, New York Attn: Dr. F. S. Cooper	1

Code	Organization	No. of Copies
I 70	Melpar, Inc. 3000 Arlington Boulevard Falls Church, Virginia Attn: Dorothy A. Allen, Librarian	1
I 90	Litton Systems, Inc. 221 Crescent Street Waltham 54, Mass. Attn: Dr. George Sebestyan	1
I 98	Signals Research & Development Establishment Christchurch, Hants, England Attn: Walter Lawrence	1
I 104	Sylvania Electric Products, Inc. 100 First Avenue Waltham 54, Mass Attn: Charles A. Thornhill, Report Librarian Waltham Laboratories Library	1
I 140	Autonetics Division North American Aviation Whittier, California Attn: Dr. J. D. Bledsoe	1
I 186	Westinghouse Electric Corp. Electronics Division 2519 Wilkins Avenue Baltimore 3, Maryland Attn: G. H. McAdie	1
I 192	Bhimani Research Associates 1838 Massachusetts Avenue Lexington, Mass. Attn: Dr. B. V. Bhimani	1
I 193	National Cash Register Attn: Mr. Klaus Otten [illegible]	1
I 203	Philco Corporation Communications & Electronics Division Attn: R. D. McMichael 4700 Wissahickon Avenue Philadelphia 44, Pa.	1

Code	Organization	No. of Copies
I 266	ITT Federal Laboratories Technical Library 500 Washington Avenue Nutley 10, New Jersey	1
I 301	Sylvania Electric Products, Inc. Applied Research Laboratory Sylvan Road, Waltham, Mass. Attn: Mr. Harold Manley	1
I 347	Philco Corporation Advanced Communications Engineering Dept. Philadelphia 44, Penna. Attn: R. W. Steele	1
I 395	Melpar, Inc. 3000 Arlington Boulevard Falls Church, Virginia Attn: Contract Administration Department	1
I 398	ITT Communications Systems, Inc. Paramus, New Jersey Attn: Curtis M. Jansky	1
I 650	Bolt, Beranek & Newman, Inc. Attn: Dr. Karl Kryter Moulton Street Cambridge, Mass.	1
I 728	Defense Electronic Products Radio Corporation of America Attn: Willard F. Meeker, Bldg. 13-5-8 Camden 2, New Jersey	1
I 930	Hughes Communications Division P. O. Box 90904, Airport Station Los Angeles 45, California	1
M 50	AF Electronic Systems Division (ESNC) SPACECOM (Attn: 4841, Maj. Courtney) 424 Trapelo Road Waltham, 54, Mass.	1
N 15	U. S. Naval Research Laboratory Code 5418, Attn: Barry E. Schuliger Wash 25, D. C.	1

Code	Organization	No. of Copies
N 35	Commanding Officer and Director U. S. Navy Underwater Sound Laboratory Fort Trumbull, New London, Connecticut Attn: Mr. C. M. Dunn	1
N 62	Material Laboratory Library Building 291 Code 912B New York Naval Shipyard Brooklyn 1, New York	1
N 137	U. S. Naval Electronics Laboratory Attn: Dr. John C. Webster San Diego, California	1
U 7	University of California Department of Engineering Berkeley 4, California Attn: Antenna Laboratory	1
U 19	Prof. Joshua Whatmough 17 Central Street Winchester, Mass.	1
U 29	Massachusetts Institute of Technology Lincoln Laboratory Lexington, Mass. Attn: Dr. James Forgie	1
U 50	Speech Transmission Laboratory Royal Institute of Technology Stockholm 70, Sweden Attn: Dr. Gunnar Fant	1
U 102	Harvard University Technical Reports Collection Gordon McKay Library 303A Pierce Hall, Oxford Street Cambridge 38, Mass. Attn: Librarian	1
U 248	University College Phonetics Department Attn: Dr. Adrian Fourcine Gower Street London WC-1, England	1