Segmentation and Labeling of Speech: A Comparative Performance Evaluation

Henry Gilbert Goldberg

December 1975

DEPARTMENT
of
COMPUTER SCIENCE

Carnegie-Mellon University

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: AFOSR - TR - 76 - 0604	2. GOVT ACCESSION NO. / 3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): SEGMENTATION AND LABELING OF SPEECH: A Comparative Performance Evaluation	5. TYPE OF REPORT & PERIOD COVERED: Interim rept. / 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Henry Gilbert Goldberg	8. CONTRACT OR GRANT NUMBER(s): F44620-73-C-0074, ARPA Order-2466
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Carnegie-Mellon University, Computer Science Dept., Pittsburgh, PA 15213	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: AO 2466, 61101D
11. CONTROLLING OFFICE NAME AND ADDRESS: Defense Advanced Research Project Agency, 1400 Wilson Blvd, Arlington, VA 22209	12. REPORT DATE: December 1975 / NUMBER OF PAGES: 212 (214 p.)
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): Air Force Office of Scientific Research (NM), Bolling AFB, DC 20332	15. SECURITY CLASS. (of this report): UNCLASSIFIED / 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

(studies)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This thesis is a study of speech recognition at the parametric level. It attempts to evaluate and understand the relative merits of a number of alternative design choices at that level. Such a study raises issues in Artificial Intelligence, Linguistics, Acoustics, Pattern Recognition, Statistics, and Speech Understanding research. In particular, it involves an investigation of segmentation and labeling techniques, and the use of parametric representations for the acoustic signal in those techniques. Every speech recognition system employs some parametric representation and some initial signal to symbol transformation. We show the performance currently available for these initial (Continued)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

processes, and asserts that such performance is comparable to human performance. We present the relative merits of some typical parametric representations, and develop a methodology for such comparative evaluation. Simple, parameter-independent schemes for segmenting, labeling, and training are developed as well. The role of pattern classification techniques is clarified, as it relates to the initial signal to symbol transformation.

Four parametric representations have been chosen for study: a set of amplitudes and zero-crossing measurements from 5 octave filters (ZCC); a set of energy measurements from a 1/3 octave filter bank (ASA); a smoothed, short-time spectrum computed from the LPC filter (SPG); and the LPC coefficients themselves (ACS). Note that the first two involve the use of analog devices. Each method yields a set of measurements at uniform, short intervals -- a pattern. Distance functions, chosen from Pattern Classification theory, are then applied to the parameter patterns as measures of acoustic similarity.

A method for segmenting speech into isolated, acoustically consistent segments is presented. The method is fairly independent of the choice of parametric representation, since it relies upon the acoustic similarity measure as the primary evidence of acoustic change. Missing and extra segment errors are found to be as good as 4% and 19%, respectively. Significant differences in the segmentation effectiveness of the parametric representations is found. They may be ordered as follows: SPG, ACS, ASA, and ZCC. The best performance is found to be comparable to the state of the art. Little reduction in accuracy is encountered when new speakers are tested.

Labeling is accomplished by the same pattern similarity measures. However, similarity is measured between the unknown pattern and each of a set of stored templates. A clustering algorithm is presented which finds the most suitable set of templates to represent a population of patterns which correspond to a particular phonetic label. The patterns tested are those isolated by the best machine segmentation routine, hand corrected for serious errors.

Little difference is observed along the parametric representation or the classification metric dimensions, except for poorer performance for ZCC input. Each input segment is labeled as one of a set of 40 phone labels. The correct phone appears as the first choice 28% of the time. It appears in the first three choices 55% of the time. However, when a lower level, acoustic transcription is used as evaluation referent, these values increase to 42% and 65%. Even the 28% accuracy, which arises from a comparison against phonemic expectation, is acceptable performance. It is the same as or slightly better than human spectrogram reading performance in the absence of other linguistic clues.

The major contributions are as follows. 1) Simple yet effective, parameter-independent procedures for segmenting and labeling speech are developed; (2) A methodology for performance evaluation at this level is presented; (3) A number of alternative design choices are examined. 4) A better understanding is offered of the role of pattern classification techiniques in the initial signal-to-symbol analyses.

Segmentation and Labeling of Speech: A Comparative Performance Evaluation

Henry Gilbert Goldberg

December 1975

Department of Computer Science
Carnegie-Mellon University
Pittsburgh, Pennsylvania 15213

Submitted to Carnegie-Mellon University in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

This research was supported in part by the Advanced Research Projects Agency of the Office of the Secretary of Defense (Contract F44620-73-C-0074) and monitored by the Air Force Office of Scientific Research. This document has been approved for public release and sale; its distribution is unlimited.

Abstract

This thesis is a study of speech recognition at the parametric level. It attempts to evaluate and understand the relative merits of a number of alternative design choices at that level. Such a study raises issues in Artificial Intelligence, Linguistics, Acoustics, Pattern Recognition, Statistics, and Speech Understanding research. In particular, it involves an investigation of segmentation and labeling techniques, and the use of parametric representations for the acoustic signal in those techniques. Every speech recognition system employs some parametric representation and some initial signal to symbol transformation. We show the performance currently available for these initial processes, and assert that such performance is comparable to human performance. We present the relative merits of some typical parametric representations, and develop a methodology for such comparative evaluation. Simple, parameter-independent schemes for segmenting, labeling, and training are developed as well. The role of pattern classification techniques is clarified, as it relates to the initial signal to symbol transformation.

Four parametric representations have been chosen for study: a set of amplitudes and zero-crossing measurements from 5 octave filters (ZCC); a set of energy measurements from a 1/3 octave filter bank (ASA); a smoothed, short-time spectrum computed from the LPC filter (SPG); and the LPC coefficients themselves (ACS). Note that the first two involve the use of analog devices. Each method yields a set of measurements at uniform, short intervals -- a pattern. Distance functions, chosen from Pattern Classification theory, are then applied to the parameter patterns as measures of acoustic similarity.

A method for segmenting speech into isolated, acoustically consistent segments is presented. The method is fairly independent of the choice of parametric representation, since it relies upon the acoustic similarity measure as the primary evidence of acoustic change. Missing and extra segment errors are found to be as good as 4% and 19%, respectively. Significant differences in the segmentation effectiveness of the parametric representations is found. They may be ordered as follows: SPG, ACS, ASA, and ZCC. The best performance is found to be comparable to the state of the art. Little reduction in accuracy is encountered when new speakers are tested.

Labeling is accomplished by the same pattern similarity measures. However, similarity is measured between the unknown pattern and each of a set of stored templates. A clustering algorithm is presented which finds the most suitable set of templates to represent a population of patterns which correspond to a particular phonetic label. The patterns tested are those isolated by the best machine segmentation routine, hand corrected for serious errors.

Little difference is observed along the parametric representation or the classification metric dimensions, except for poorer performance for ZCC input. Each input segment is labeled as one of a set of 40 phone labels. The correct phone appears as the first choice 28% of the time. It appears in the first three choices 55% of the time. However, when a lower level, acoustic transcription is used as evaluation referent, these values increase to 42% and 65%. Even the 28% accuracy, which arises from a comparison

against phonemic expectation, is acceptable performance. It is the same as or slightly better than human spectrogram reading performance in the absence of other linguistic clues.

The major contributions are as follows. 1) Simple yet effective, parameter-independent procedures for segmenting and labeling speech are developed. 2) A methodology for performance evaluation at this level is presented. 3) A number of alternative design choices are examined. 4) A better understanding is offered of the role of pattern classification techiniques in the initial signal-to-symbol analyses.

Acknowledgements

I would like to acknowledge the aid, both scientific and moral, which has been so generously offered me. My advisor, Professor Raj Reddy, has been a true and patient friend as well as a challenging mentor. The members of my thesis committee, Professors Samuel Fuller and Paul Shaman, and Doctor Lee Erman, have provided careful and expert criticism -- always constructively. I would like to thank each of them for his support. In addition, I am grateful to Professor Allen Newell for pointing out the applicability of the signal detection model to this work and for significantly contributing to the atmosphere and resources at Carnegie-Mellon University which are so conducive to scientific research. For that environment, I am indebted to the entire Computer Science Department. I would like to thank my friends and colleagues, some of whom were friendly competitors for computing resources. If I have succeeded, it is because they let me stand on their toes. I am particularly grateful to Doctor Mario Barbacci who acted as foil to my ideas on many occasions and to my good friend, Caragos Lapin, who never asked me when I would finish. Finally, I would like to thank my wife, Carol, for her moral support. This thesis and our marriage were begun at about the same time. I can have no fonder hope than that the former be forgotten long before the latter.

CONTENTS

Chapter 1

Background and Problem Statement

In recent years, a renewed attack has been made on the problem of input of human speech to computers. [New71,Red75b] This dissertation is particularly concerned with one component of this problem -- the initial analysis of the acoustic input. A great deal of our understanding of this problem has come from areas such as linguistics, physiology, acoustics, and psychology. Computer science, and in particular artificial intelligence, has played a catalytic role in drawing together knowledge from diverse sources into workable structures. Common to all these structures is a component which deals with the acoustic input in some parametric form. From that component we expect an initial isolation or identification of the information borne by the acoustic signal. In this thesis we focus on this essential element, its inherent problems, the issues involved in its implementation, and its role in a total system.

1.1 Introduction

The basic vehicle for this research is the problem of choosing a parametric representation for the acoustic signal which is to be input to a speech understanding system. The choice must ultimately be made by the individual system designer for there is, as yet, no one clearly superior parametric representation that serves the variety of purposes of segmentation, phonetic analysis, prosodics, *etc.* which are needed to understand general continuous speech. Up to this point, the prospective system builder has made the choice in an *ad hoc* manner. Either certain hardware was already available, or the necessities of cost and/or time prevailed. In other cases, representations were based upon traditional methods. In those cases where a parametric representation was developed from first principles, those principles have consisted of limited empirical studies, often influenced by the element of human speech understanding ability, or they have been based upon simplified assumptions about the physical or stochastic nature of human

speech. In short, we are faced with a number of different methods for extracting acoustic parameters from the speech signal. All are based upon reasonable, but not complete, understanding of the nature of the speech signal. Some make trade-offs with speed and cost which may not be suitable or necessary. Many have been employed in speech understanding systems of varying complexity and success. Some can be shown to support re-synthesis of speech. But very few have been comparatively examined in the light of their eventual use in a total system. (See [Fla72] for a survey of speech analysis and synthesis techniques.)

In order to make the comparisons so that they will be useful to the speech system designer, three problems must be considered. 1) The role of the acoustic information and knowledge about acoustic-phonetics, in the context of the entire system, should be understood. 2) The method by which the acoustic parameters are analyzed -- the recognition scheme -- should be chosen with care. 3) The performance statistics must be designed to convey sufficient information about the abilities of a parametric representation to support recognition. The information is needed by the designer to predict what the choice will mean in terms of his system.

This chapter is a statement of the problem to be attacked. As such, it must survey the terrain before proceeding. In the following section, we will discuss those aspects of speech understanding systems which seem to be relevant to the question of system use of the acoustic parameters. Section 1.3 is a look at the uses to which the acoustic knowledge itself is put -- what kind of processing will be needed depends upon what kind of information about an utterance is required at the acoustic level. Section 1.4 is a survey of the available methods for extracting parametric representations of a speech signal. And the final section states the specific problem in terms of the limitations, assumptions, and performance dimensions chosen for study.

In succeeding chapters, we will present a very brief survey of pattern classification ideas and methods, chapter 2, since these concepts are so basic to the type of analysis done at the parametric level of speech understanding systems. Chapter 3 will discuss

aspects of the pattern classification problem particularly relevent to speech recognition. In addition, a brief survey of the acoustic/phonetic processing of a number of current systems is included in chapter 4 to provide some context for this work and for other results in the area, as well as to provide some idea of the currently available technology and performance.

Chapters 5 and 7 will present and discuss the methods for segmentation and labeling, respectively, used in this research. Chapters 6 and 8 will present the methodology for evaluation of performance and the results obtained. Finally chapter 9 is a concluding discussion which will serve to focus attention on the most important elements of this work, and will provide an appropriate overall view for evaluating results of research at the level of acoustic-parametric analysis.

1.2 Speech Understanding Systems

In this section, we will discuss Speech Understanding systems. Speech Understanding invloves the input of a speech utterance, the extraction of relevent linguistic information from the acoustic input, and the decoding of that information into some meaningful construct. A distinction is often made between Speech Recognition -- the process of extracting information by the use of knowledge about speech -- and Speech Understanding -- where knowledge about the meaning of the utterance may be used to decode it. The purpose of the discussion is to provide enough of an overall picture of these systems that the acoustic analysis problem can be seen in perspective to the total problem. Since there is little difference for our purposes between these two types of speech system, we will use the terms interchangeably.

At first glance, the problem of understanding the role played in speech understanding systems by acoustic parameters might seem to be insurmountable. Clearly different systems will use their acoustic knowledge sources differently. Their other parts will interact with each other in very different fashions. Errors fatal to some systems might be easily corrected by others. However, this apparent lack of any unifying model of a

speech understanding system is not total. One may assume some structure and limitations for the purpose of studying systems to be developed now and in the near future. There is no clear model of what a speech understander should look like except the human model, which is not describable to any great extent as yet. The information we do have about human speech is structured into well defined theories or levels, and this structure can tell a lot about the form that speech systems will take and the role that acoustic (parametric) knowledge and analysis will play in them. The variations among systems become, in this view, more questions of degree than of essential differences. How much weight does one give to semantically based inferences about the utterance? How powerful a model of the speaker is available? *etc.* The answers to such questions of relative merit of the various types of knowledge about speech and speakers gives flesh to the skeleton structure of the different levels. Then a control structure for handling interactions among the levels is imposed so that errors can be detected quickly, work can be shared and efficiently performed, and the knowledge source most likely to succeed can be invoked in any situation.

1.2.1 Sources of Knowledge

In their report on speech understanding systems, Newell *et al.* [New71] point out the relevance of the levels commonly accepted by linguists and phoneticians to questions of system structure and control. It is important to note that every system developed to date has a number of internal representations of the input utterance. These representations correspond to the levels of discourse in speech science such as the acoustic, phonetic, lexical, syntactic, and semantic. Working at various levels are sources of knowledge about speech which serve to translate from one representation to another. In these processes, such recognition activities as search, classification, error correction, hypothesizing, and verifying may occur. (see figure 1.1) A source of knowledge at the word level, for example, may initiate a lexical search to convert a phonetic sequence into a word. Or it may be used to generate a sequence of phones to be verified or matched against the input at any of a number of lower levels.

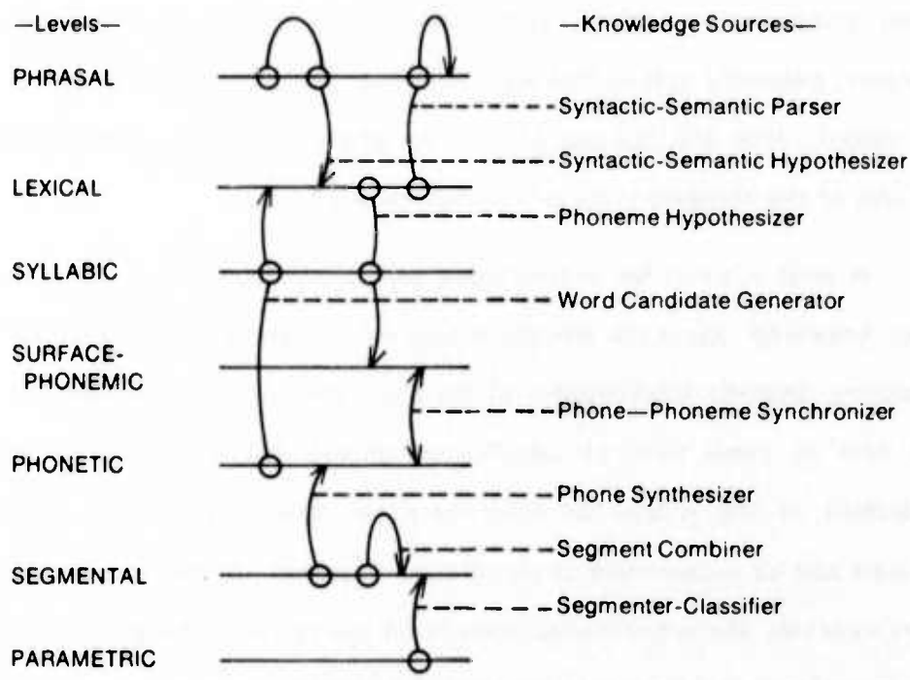


Figure 1.1: Levels and Knowledge Sources in HSII

Using the common data representations and speech knowledge of traditional theory to relate speech understanding systems to one another, one can hope to draw some conclusions (albeit general ones) about the role of acoustic knowledge sources and data in an entire system. People often conceptualize the structure of speech understanding knowledge application as one of a linear flow through the levels. Either bottom-up or top-down strategies of search allow decisions (and errors) to be transmitted and transduced through the levels in a rather straightforward manner. However, the interactions among levels may, in general, be complex. One cannot assume any particular form for the control flow of such systems, but we will briefly discuss below a number of forms that have been applied to speech understanding systems.

Two data representations common to many systems are the acoustic parameters and a phonetic-like transcription. The knowledge sources that we are investigating in this

dissertation are partly responsible for translating from the former to the latter.† Some limited word recognition systems have shown great success bypassing the phonetic transcription and recognizing words directly from the acoustic input parameters. It is, however, generally agreed that such techniques fail with connected speech for a number of reasons. (For one, the lack of word boundaries will cause an exponential increase in the size of the recognition pattern storage required.)

In most systems for understanding general continuous speech, the processes which apply knowledge about the acoustic and phonetic nature of speech gestures to the task of producing phonetic transcriptions of the signal play a very important role. Essential to this task is some form of classification scheme and some process for segmenting, regardless of the manner in which these two processes interact. Segmentation may preceed and be independent of classification (labeling). A label may be chosen at regular short intervals and segmentation procede on the resultant string. Or the two processes may operate on the same data and interact to support or reject each other's decision. In any case, a parametric representation which does not reflect a particular acoustic cue of segment boundary will produce segmentation errors, and one which maps different acoustic realizations of phones into the same parameters with produce labeling errors.

1.2.2 Some Control Structures

Although there is no one structure for interactions among the knowledge sources of a speech understanding system, there are a few paradigms of such interactions which have been proposed and applied to working systems. All of these paradigms deal with information about the input utterance represented internally at a number of levels in some incomplete, possibly errorful, data structure.

Systems organized to interact in a linear manner tend to be susceptible to error propagation through the levels. However, subsystems of a number of speech recognizers

† Other sources of knowledge, concerned with phonetics, coarticulation, and stress for example, are needed to deal with truly general speech. To deal with this straightforward translation, it appears that classification based on acoustic patterns alone is not powerful enough.

do obey linear control flow where a sequence of separate sources of knowledge each act upon the previous one's output and feedback is initiated only from certain levels. An example is in Hearsay I where broad classification, segmentation, fine classification, and lexical search are linearly invoked, but feedback only results from higher levels initiating a different lexical search. [Erm74b] This is a case where everything that can be done in the general area of acoustic analysis of the utterance is done immediately. Thus, there is no purpose to invoking any inter-level paths other than the straightforward one that reduces the representation to the highest level data structure used in the system.

An early paradigm for speech recognition, suggested by Halle and Stevens [Hal62], is Analysis-by-Synthesis. A representation of the input is postulated at some level and the sources of knowledge are used to create a corresponding representation at another lower level to be compared with the input. Some measures of closeness of the two representations at the lower level are used to decide upon the "truth" of the higher level assertion. Again, a linear system structure is likely to be used here since the point at which feedback is initiated is at the low level comparison, after a sequence of transformations of the represented synthetic utterance. Analysis-by-Synthesis can also be applied in subsystems where the rules are available in a powerful but generative form, and the size of the search for the correct representation to synthesize is not excessively large. [Kla75]

The Hearsay system paradigm of Hypothesize and Test [Red73] is similar to, but more general than, Analysis-by-Synthesis. The test need not be a comparison of two structures at the same level. In fact, the test will most often be constructed to compare only those parts of the representation which may feasibly differ in a teleological sense (in the sense that they might lead to different results at the higher levels). Flow of control among the levels is much less constrained, and consequently the interactions are more complex.

Various parts of speech understanding systems may be treated as heuristic searches in the sense that a universe of feasible solutions (interpretations for the input

representation to the subsystem) is being searched by application of specialized rules dependent upon the current point in the universe being investigated. Knowledge sources that allow representations at one level to be recognized only as legal representations at another higher level are the operators that traverse the universe of solutions. Heuristics for applying the operators may be explicit in some scoring mechanism or implicit in the knowledge sources. (E.g., when a key syntactic element is discovered, it is reasonable to generate the surrounding modifiers or function words.)

Dynamic programming techniques have been successfully applied to simple, powerful systems for word or short phrase recognition. [Ita75, Fu68, Ich73, Whl75] Usually a single source of knowledge -- an acoustic classifying scheme -- is used within the dynamic programming algorithm to find the best fit among a number of stored templates. The dynamic program provides the ability to adjust time durations of the various segments to a limited degree without explicitly segmenting. This is a very powerful technique for short utterances from a limited set and may be used as a component within a speech understanding system.

Baker [BakJK75b] presents the Hidden Markov process as a model for recognition at each of a number of levels, implemented as a dynamic program. Flow of control in his system is handled by the probabilistic model itself. An underlying representation of each level is hypothesized as a Markov sequence which best fits the observed representation. At each level, elements of the lower level representation may stand for realizations of elements of the representation in view. These latter are connected in a standard Markov chain. The probability of a realization is a combination of the underlying (higher level) chain's probability and the individual realization probabilities. The translation of the underlying sequence to that of a higher level is much simpler since it is more highly constrained than the observed representation.

Our purposes in briefly discussing these models of system interaction are twofold. First, one can see that, inherent in all the systems thus far developed, there is the action of translating a piece of one level's data structure into that of another. At the acoustic

level, this almost always means some form of classification of a short interval of the acoustic representation into one of a number of phone-like classes. Measuring the performance of such an action for the different acoustic representation schemes will, therefore, provide information relevant to the performance of the vast majority of speech understanding systems. The second purpose is to point out the feasibility of using models of performance of knowledge sources in an analysis of the entire system's performance. Although the control paradigms affect the order of applying the different knowledge sources and the amounts of effort wasted on false paths or bad hypotheses, the progression of the correct representation through the levels is universal. Some piece of the input signal will have been transformed by a sequence of classifications into either a phonetic sequence, a word, or a phrase element. In continuous speech systems, further transformations will have eventually carried these elements to a single semantic or task related construct. While the entire system analysis may require simulation, if no analytic model is available, the individual knowledge sources are separable and their effects on system performance are separable.

1.2.3 Human Performance

A great deal has been written about all aspects of human perception of speech, and we cannot even survey what is known or postulated about the structure and interactions of knowledge within the human speech understanding system. However, the existence of human speech perception under all manner of difficulties and limitations does point to ways of analyzing individual knowledge sources for their role in the total picture.

Experiments in perception of words under noisy conditions have quantified to some extent the role of semantic support in disambiguating errorful inputs. [Bru56] In a like manner, errors in perception are correlated with ungrammaticality to measure the role of syntax. An experiment involving unfamiliar languages [Sho74a] has shown some interesting results as far as the accuracy of human phonetic recognition is concerned. In this last, expert phoneticians are presented with utterances in a number of languages whose words, syntax, and phonology are totally unfamiliar (Turkish, Cantonese, Swedish,

etc.). They are asked to produce as accurate a phonetic transcription as possible from listening to the recorded utterances or from observing graphical displays such as sound spectrograms or oscillograms. Very briefly, using auditory input the subjects achieved about 50% recognition at the phonetic level, with a chuoice of about 50 phone-like labels. With oscillogram or spectrogram input only, accuracy was about 25%. The results indicate that the acoustic knowledge source in human perception is not much better than the best machine procedures currently available. The human perceiver is much more adaptable and more robust over a wide range of conditions than the machine at this level. But it seems entirely likely that present techniques could, under favorable conditions, perform the foreign language experiment as well as the human subjects.

There is disagreement on whether higher level knowledge or low level recognition techniques are the bottleneck at this point. It is our opinion that there is much more to gained from improvements to higher level knowledge sources. This does not stop us from continuing to improve the acoustic level procedures available, until they are as good or better than human ability, but it does point out the need for a clear understanding of their performance characteristics. With such an understanding, system design efforts may be best directed, and the results of improved higher levels will be recognized.

1.2.4 Summary

We have presented a picture of speech understanding systems as collections of separable sources of knowledge, with representations of the speech signal occuring at a variety of levels. The manner of interaction among these knowledge sources is of varying importance in analyzing their performance. Our view is that the acoustic level processes are particularly easy to separate.

1.3 Acoustic Level

This section will discuss the role that acoustic knowledge can play in a speech understanding system and the types of decisions that can be supported by it. The parametric representation of the utterance to be understood may be considered as the raw input of the system. Most higher level knowledge is not expressed in terms of this representation. For this reason as well as the quantity of data that is input, some serious reduction of the amount of data and some translation into another representation are the primary requirements of this level. In addition, the system needs a reasonably powerful way to begin its search for a solution. In some situations, semantics or syntax may be able to provide such a handle, but often one must rely upon the acoustic input to make an initial hypothesis from which the rest of the system may proceed. These three actions -- *data reduction*, *translation*, and *hypothesis generation* are the most common uses for acoustic level analysis in speech understanding systems.

The two types of processing that are typically applied are segmentation of the utterance into quasi-phonetic segments and labeling of those segments with information interpretable by higher levels -- usually indentifying phone-like sounds. Although the production of an actual phonetic transcription might involve a number of sources of knowledge concerned with coarticulation, phonetics, prosodics, *etc.*, an initial translation into a sequence of acoustically separate segments and their classification into types of speech sounds can provide a reasonable first approximation at a transcription. [GolH74] It is our contention that a simple segmentation and labeling scheme can be used in this comparison study. That is not to say that the limits of acoustic knowledge sources are such simple schemes, but rather that these two basic processes are elementary processes that more complex algorithms will depend upon. It is also an assertion that the primary role of acoustic level analysis is satisfied by these two processes. The following brief discussions should give a better idea of both the kind of processing to be done with parametric representations of speech and the roles that the results of such processing play in the whole system.

1.3.1 Segmentation

The segmentation process is conceptually simple -- to find the boundaries in time between the different sounds that make up an utterance. The difficulty seems to lie in defining what is meant by "different sounds". In a phonetic or phonemic segmentation, some segments are essentially steady state in their acoustic characteristics, others are continuously varying or transitionary in nature, and some are composites of two or three sounds of either type. An acoustic segmentation, on the other hand, separates the input into portions within which the acoustic character is consistent. Transitionary sounds will still present a problem. For example (see figure 1.2), the sound /l/ displays a time varying resonant structure, as does the initial portion of a vowel following a /g/ or the middle portion of a diphthong. Yet only in the first case would everyone agree that a separate sound must be identified and set apart from its neighbors. Clearly, the fineness of resolution to which one requires segmentation be done depends upon the final uses one has for a machine transcription of the utterance. If differentiation of words is done by crude identification of consonants and careful analysis of the most stressed vowel, for example, then segmentation should be biased towards identifying the long steady state portions as single segments, even at the cost of losing some consonant segments. If consonants are identified by their coarticulative effects upon neighboring phones, transitionary portions become very crucial and must be located. In general, the commonly accepted phonemes of English (or whatever language is being spoken) give an idea of the degree of resolution needed for most analyses. If the segmenter can separate those portions of the signal most likely to be associated with the phonemes that make up the utterance, then small variations in how diphthongs, plosives, etc. are treated are not critical. If the speech understanding system relies upon a set of labels for sounds that are considerably different from the phonemes, the segmentation must be able to separate

those sounds robustly.†

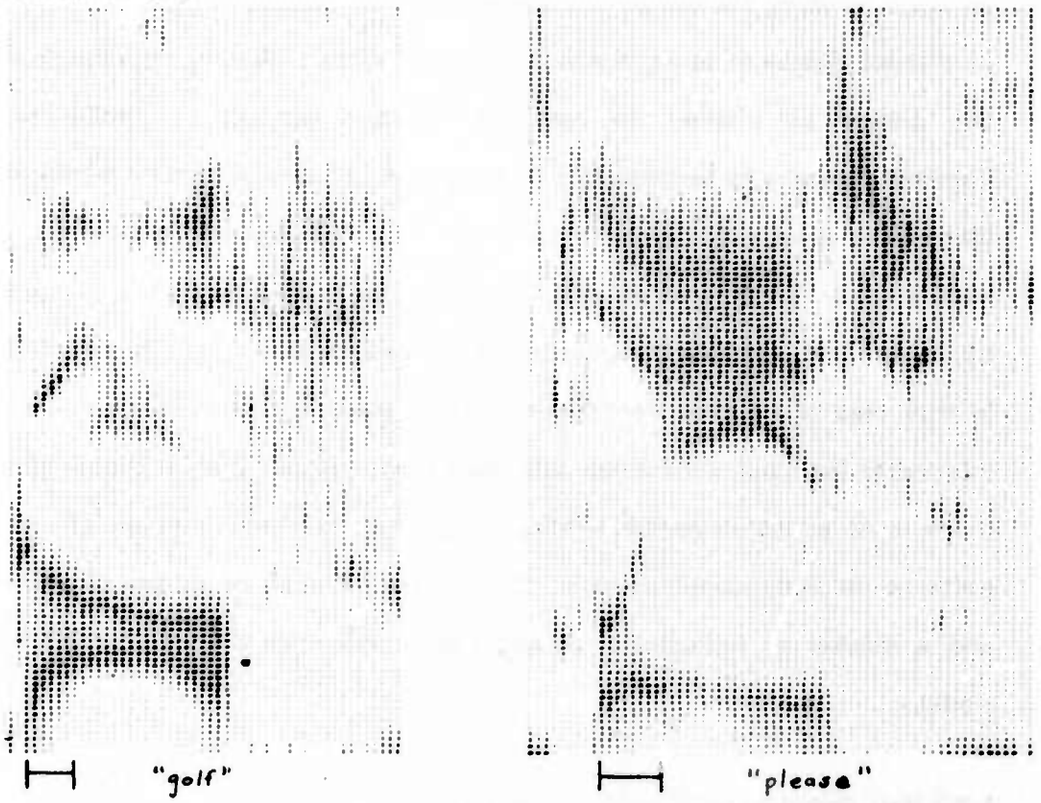


Figure 1.2: Some Time-Varying Segments

1.3.2 Labeling

The labeling process is the central pattern recognition process at the acoustic level. The parametric representation of an input segment of speech is labeled with an indicator of the information it is deemed to be carrying. Until some such labeling is accomplished, the sequence of segments may be any sequence at all. Thus 10 segments, each of which may be any of 30 types of speech sound, represent 30^{10} possible transcriptions of the input. The labeling process, by reducing the 30 choices to, say, 3, can reduce the search by a factor of 10^{10}. This possibility of reducing the exponential search size is due to the fact that the acoustic labeling and segmenting are applied first, when little else is known about the utterance, and that the vast majority of representations at this level are illegal at higher levels and would never have been produced by the speaker in the first place.

† One segmentation process in Hearsay I picks out voiced, fricated, and silent segments only. A later process may subdivide these segments upon more detailed analysis.

The issue of what are the labels that are to be placed upon the input utterance is an issue involving the design of a number of levels. Whether labels are to be considered as distinct classes or as regions in a continuous space of labeling information is central to the choice of whether to recognize phonetic features or phone-like gestures. Segmentation may be accomplished by labeling at regular short intervals and then marking boundaries at maximal changes in the labels. In such a situation, the label set must reflect such a goal. The dynamic programming model that is used as a word recognition system by Itakura and others labels an entire short utterance as a word. The primitive operation in that case is a pattern recognition measure which determines how close a fit a short interval in the input word makes with the stored template. Even in such a system, where there is no actual phone-like labeling being done, the primitive action of comparing two patterns for likely identity is basic. Chapter 2 will discuss the pattern classification model and a number of methods for solving simple recognition problems within that general model.

1.3.3 Data Reduction

A typical digitized signal contains at least 10K samples per second, where each sample should be at least 9 bits, probably more.† The parameters extracted from the signal may reduce this data rate considerably. Spectrograms offer no reduction per se, although the locations and amplitudes of spectral peaks (formant tracking) represent approximately an order of magnitude saving. Typical analog filter banks, digitized every 10ms., offer the same order of magnitude reduction. However one is still faced with perhaps 10K bits per second, and only the most straightforward analysis can keep up with such a data rate. Thus, an important role of the acoustic analysis level is the reduction of the input data rate to an amount manageable by the higher levels, where interactions, backtracking, and more complex analysis will preclude large, redundant data representations. Merely labeling each 10ms. interval with one of a set of about 50 labels

† In fact, 16-bit accuracy or a floating point scheme is needed. In dealing with 9-bit data, our experience has been that not enough dynamic range is available. Either stressed vowels are clipped, or unstressed nasals lack any waveform structure.

reduces the rate to 600 bits/second. Further reduction is available by forming segments with one label for a longer duration of signal (typically from 10 to 200ms., usually 50 to 100ms.). However, this latter saving may be spent on multiple labels, rating schemes, and certain special parameters, such as overall amplitude, which may be useful to other knowledge sources. The data in its new form is not only more compact, but also much less redundant.

1.3.4 Translation

It has often been pointed out that a problem in applying much of the codified knowledge about speech is that it exists in terms of generative rather than analytic rules. However, another serious problem in applying such knowledge is that the rules are written in terms of very different primitives. For example, syntax is often understood in terms of lexemes -- words or endings of words; coarticulation rules are in terms of phones or other perceptual features. The difficulty is that making a clear and universal correlation between such elements and another representation, such as the acoustic parameters, is not possible. (That is what speech recognition is all about.) Clearly, some initial translation must be made from the acoustic parameters to some other representation better suited to application of these rules. Most system designers have chosen the new representation to be some form of phonetic label† although this need not be the case. The new representation may consist of entirely heuristic elements, pseudo-phonemes, or even, as in some word recognition systems [Ich73, Ita75] entire words. The latter case is one where no other knowledge is applied to the utterance except the acoustic matching in the context of a dynamically adjusted time scale. The point to be made is that the role of acoustic level translation is determined by the data structures of the other sources of knowledge.

---

† The term "phonetic" carries implications of more human perception orientation than is usually available. Indeed, one could argue that machine labels merely represent classes of sounds with certain acoustic characteristics. They are no more phonetic or phonemic in nature that any other sounds. However, it is usually the goal in defining these classes to pick sounds whose acoustic characteristics correlate highly with phonetic or even phonemic information. In this sense, machine labels can carry both acoustic and phonetic information.

Most systems have adopted a phonetic data representation at some level. Even if a system has no such representation, translation still occurs in some form, from the parametric representation to some other representation.

1.3.5 Hypothesis Creation

In a system which attempts to develop a partial representation of the utterance at higher levels, the key to successful recognition is often the ability to create a "handle" early in the process. Figure 1.3 shows an example of hypotheses created in Hearsay II. Some phoneme, word, or phrase is recognized with high confidence, and the search spaces of a number of different levels are significantly reduced. In addition, many rules of both generative and analytic nature deal with elements in some limited context, so that inference can only be made when some such context is available. It is, therefore, an important role of the acoustic knowledge sources to provide initial hypotheses about the utterance from which inferences may be carried forward, verified, or altered. Some system structures, such as Analysis-by-Synthesis, do not proceed in this fashion. Rather, the entire utterance is generated or stored as a template and a complete test is made. Most implementations of such methods are restricted to particular levels with more flexible overall control of the system; then the results of such tests are used on only limited portions of the utterance. It is generally accepted that systems (in order to be robust in the presence of errors) will require the ability to create hypothetical recognitions and to alter them as new information is discovered. Therefore, the acoustic level results will have to be viewed as an important source of such initial hypotheses or at least as the first source of verification decisions. Issues such as: how confident one can be in a particular piece of the result, how often a really solid handle is found, and how errors will affect the usefulness of the results as hypotheses for the rest of the system, become important to the analysis of performance and the prediction of merit to a working system of an acoustic level recognition scheme.

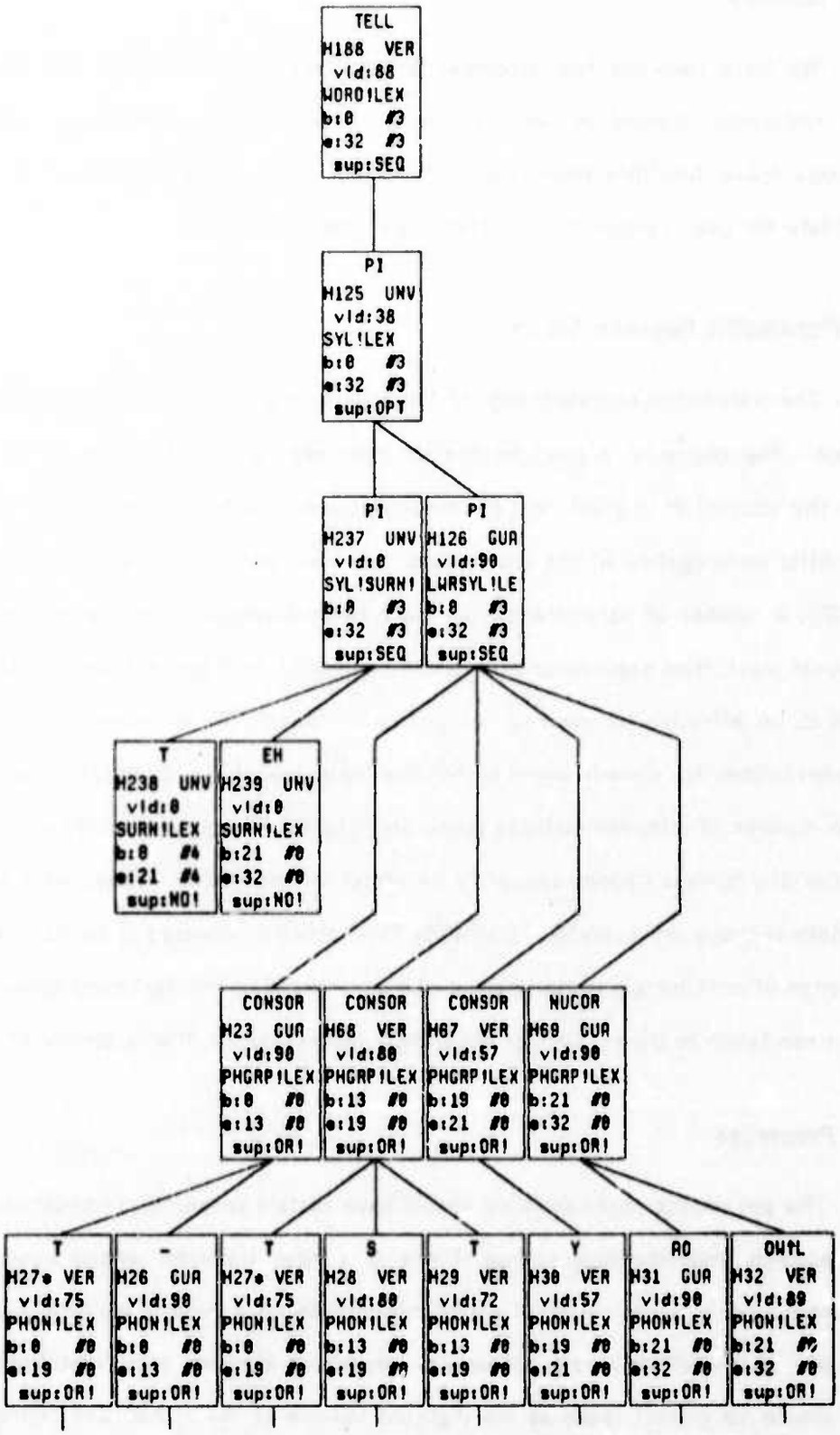


Figure 1.3: Hypotheses in HSII

1.3.6 Summary

We have seen the two processes of Segmentation and Labeling and their roles in data reduction, translation, and hypothesis creation. Any knowledge source which provides these functions from acoustic to higher level representations is a satisfactory candidate for use in a speech understanding system.

1.4 Parametric Representations

The parametric representation of the acoustic signal is the basic input to the entire system. The choice of a good method for representing the utterance at this level has been the subject of a great deal of research, conjecture, and rationalizing. Even though very little investigation of the choice itself has been done (see Ichikawa for an example [Ich73]), a number of parameterizations have been developed from theoretical models of the vocal tract, from experience with human perception, or from experience with heuristics found to be effective for machine recognition of speech. An extensive survey of all the representations for speech would be beyond the scope of this dissertation, both because of the number of different methods (some only slightly different from others) and because only certain representations appear to be useful for recognition. Reasonably current and complete surveys are available. [SchRW75] This section is intended to be more a sketch of the range of possible parameterizations, and a statement of the significant approaches that have been taken to the problem of designing a representation, than a survey of the field.

1.4.1 Properties

The parametric representation should have certain properties in order to be useful to a speech understanding system. There is a clear trade-off among cost, either of implementation in hardware or of digital computation on a general purpose machine, and flexibility and small data rate. However, somewhere between representations that are very simple to extract (such as the digitized version of the signal) and representations

that are very flexible and parsimonious (an extreme example being a sub-phonemic transcription), is the parametrization best suited to each system and its resources. In addition to properties relating to cost, size of representation, and flexibility, the representation should be robust in the sense of causing the least fatal errors possible. This is a teleological property, since the seriousness of errors is only determined after the entire system is applied to the acoustic parameters. A major result of this research is intended to be a better idea of the relationship of phonetic information and the various parametric representations under investigation. In one sense, much of this question reduces to understanding what regions of the space of representations of short speech segments correlate well with useful information in the utterance, and what regions are likely to cause confusions because of their "nearness" in the space to very different information elements' patterns. In short, one hopes to find a representation which preserves the acoustic correlates of higher level information, is robust in those correlations, reduces redundant information in the signal, and is reasonably simple to extract from the raw signal. These may not all be possible at one time, or to the degree desired, but they should be considered in selecting a parametric representation.

1.4.2 Simple Parameters

Given the digitized version of an analog signal as input, there are a number of simple yet powerful measurements which can be made on the signal. Within a short time interval† where the signal is assumed stationary, the peak to peak amplitude, the positive and negative peak amplitudes, the period between major peaks, and the number of zero crossings in both or either direction may all be extracted. The pitch period, energy, voiced-unvoiced feature, and the amount of high frequency micro-structure on the waveform may all be estimated with these parameters. In particular, Baker [BakJM75] shows that a single event, the zero-up-crossing, when parametrized by the period between events and the peak amplitudes in that period gives very good information for

† The usual length of this interval is from 6ms. to 15ms. Clearly the longer the interval, the great the information reduction. Most speech gestures take longer than 10ms. to complete. Only very short burst phenomena might be lost.

segmenting and identifying stop consonants. With other measures, such as a sine-fit to measure micro-structure on the waveform, the parameters can support a general phonetic segmentation and labeling scheme. Reddy [Red68] showed that simple measurements made upon the signal and its high frequency component separately, could alone support a reasonable acoustic segmentation. Finally, even simpler measures can give useful information. Schafer and Rabiner point out the usefulness of the deltas in adaptive delta modulation schemes for detecting silence-speech boundaries. [SchRW75]

1.4.3 Spectral Analysis

A great deal of phonetic information is known to be encoded in the various frequency components of the speech signal. One often wants to separate the components of the signal according to their information content. This usually means, for speech, a transformation into the frequency domain, or some separation of the various frequency components of the waveform.

1.4.3.1 Filter Arrays

The simple measurements mentioned above may be coupled with pre-processing by analog or digital filter arrays to produce a number of signals in parallel. Besides straightforward signal enhancement by bandpass filtering to reduce AC line noise, digitization aliasing, etc., there is bandpass filtering for the purpose of isolating separate information-bearing elements of the acoustic signal. The number and bandwidth of these filters is the subject of much discussion. How well do they correspond to the formants? How costly is the array of filters to build and to digitize (in money and processing time)? The Hearsay I system uses five bandpass filters of one octave width from 200 to 6400Hz. and peak to peak amplitude and zero crossing counts on each band and the unfiltered signal. (see figure 1.4) These 12 parameters are extracted every 10ms. and used in a simple pattern classification scheme for the basic acoustic level knowledge source. We are presently experimenting with a set of 25 narrow bandpass filters which span the range of 63 to 16KHz. with ten filters per decade. (figure 1.5) Many other researchers have used

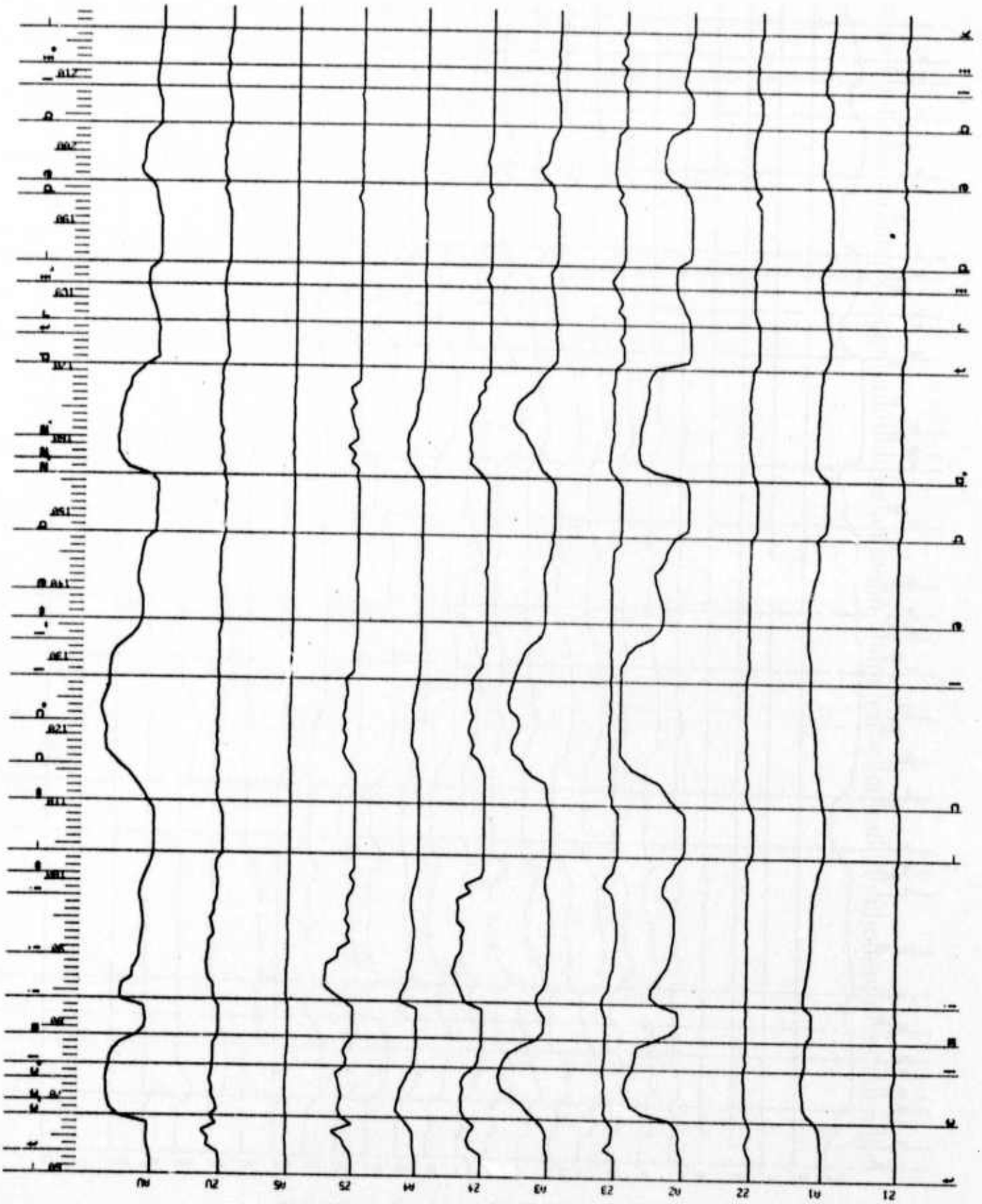


Figure 1.4: ZCC Parametric Representation

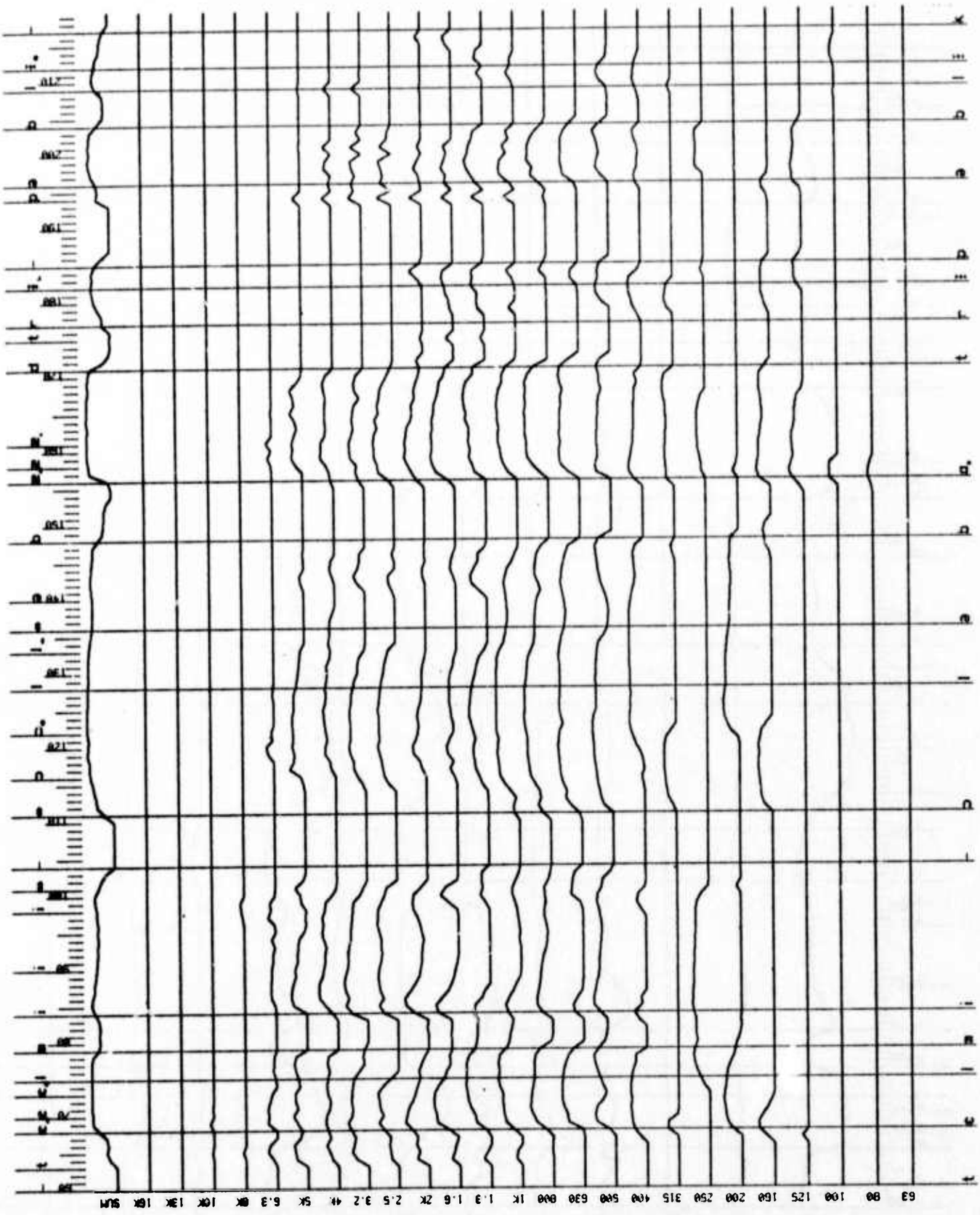


Figure 1.5: ASA Parametric Representation

arrays of filters in similar fashion to estimate a spectral analysis of the signal. These two sets of filters are fairly representative of the types and numbers of such filter arrays in current use.

1.4.3.2 Transforms

In addition to arrays of filters, whether analog or digital, the techniques of the Fast Fourier Transform may be used to calculate the frequency domain transform of a digitized speech signal quite efficiently. [Coc67] Schafer and Rabiner [SchRW75] give typical results of FFTs of speech, and discuss the various parameters of the algorithm, length of window, shape of windowing function, if any, the kind of frequency resolution obtainable, *etc.* The short time spectrum may be used to detect pitch fairly well since peaks appear in the spectrogram at harmonics of the fundamental pitch frequency. Other methods for pitch detection are also derived from the spectrum, such as the harmonic product spectrum. [SchRW75] A related method of analysis is sometimes called homomorphic filtering. The problem is to separate two signals which have been combined by multiplication and convolution. In speech processing, the central assumption is that the signal is such a combination of the excitation source and the vocal tract impulse response characteristics. Without going into details [Opp68] the log of the magnitude of the Fourier transform is the sum of the logs of the two contributors. The inverse transform, being a linear operation, preserves the additive combination in the result, known as the cepstrum. Because of this, the pitch signal, the excitation source, may be separated out and analyzed. The vocal tract impulse response may also be analyzed separately. This is accomplished by multiplying the cepstrum by a "cepstrum window" that only passes short-time components.

1.4.4 Linear Prediction

A number of formulations of a method based upon the prediction of a sample as a linear sum of the previous samples have been recently developed and fall under the term Linear Prediction or Linear Predictive Coding (LPC). These formulations, all introduced to the acoustic literature since 1966, represent a new application of a method in use by

statisticians and economists for a number of decades. However, the recent extensive work in this direction has served to demonstrate the usefulness of Linear Prediction to the analysis of speech -- particularly formant estimation -- and to provide the speech community with a number of algorithmic methods and the body of theory to support their use. As Schafer and Rabiner point out, the method is extremely powerful for the accuracy of the estimated speech parameters it provides as well as for the speed of computation possible.

1.4.4.1 Basic Method

The basic idea is that, within a short time interval (usually from 5 to 50 ms.) which is assumed stationary, the samples of the digitized signal may be expressed as a linear combination of the p preceding samples. The squared error is minimized and the least squared optimal coefficients for this prediction are found by solution of a system of linear equations.

Two formulations, which deal with slightly different treatments of the interval boundaries, are known as the Covariance [Ata71] and the Autocorrelation [Mar72, Ita68] methods. The Covariance method goes outside the interval for the p samples needed to predict the first through pth samples, while the Autocorrelation method assumes zero outside the interval. In the latter case, the interval must be windowed by a function that goes to zero smoothly at the boundaries to avoid introducing the characteristics of a step function. While the system of equations for the Covariance method is harder to solve†, Atal has shown that it requires fewer samples to achieve similar accuracy. The saving in terms of the cost of calculating over fewer samples may be significant. Neither method seems clearly superior to the other. Beside the original papers, an extensive comparison of these methods is available [Mak72] as well as shorter discussions. [SchRW75]

† The covariance system is solvable by Cholesky decomposition, for example, with approximately p^3 operations, while the form of the autocorrelation system is known as a Toeplitz matrix and may be solved by Levinson's method in p^2 operations.

1.4.4.2 Parameters

There are a number of types of parameters derivable from the Linear Prediction model. They all rely upon the same assumptions: stationarity over the interval, boundary and window choices, and size, p, of the prediction equation. However, they represent very different kinds of information about the speech signal.

The results of solving the linear equations are p parameters which are the coefficients of the predictor, or, as sometimes formulated, the coefficients of an inverse filter which can reduce the signal to noise. Itakura further processes them to remove any correlation between the ith parameter and the remaining p-i parameters. These are called the Partial Correlation Coefficients (Parcor) and have been shown to be an efficient representation for analysis and re-synthesis of speech. [Ita70, Ita68]

In actual use for speech recognition, these parameters seem to be deficient or, at best, not robust enough for simple classification algorithms. Ichikawa *et.al.* [Ich73] point out that the parcor parameters must be smoothed to achieve a reasonable recognition performance, and they still are inferior to the spectrum envelope. However, Itakura [Ita75] has developed a decision procedure from the probabilistic model of the signal used in his LPC derivations, and has shown that the predictor coefficients can be used effectively for recognition of speech.

By far, the most popular use of linear prediction is in producing estimates of the short-time spectrum envelope. The Fourier transform (using a pruned FFT) of the linear predictor impulse response, just the coefficients themselves, results in a smoothed spectrum envelope of the vocal tract response with the effects of the excitation source removed. (see figure 1.6) It is, in fact, very similar to the results of cepstrum windowing. These spectral estimates are quite accurate in locating the peak frequencies (a good guess at the formants). These locations in frequency can be derived directly from the solution of the filter transfer function, but the FFT is so fast, especially pruned for only p non-zero

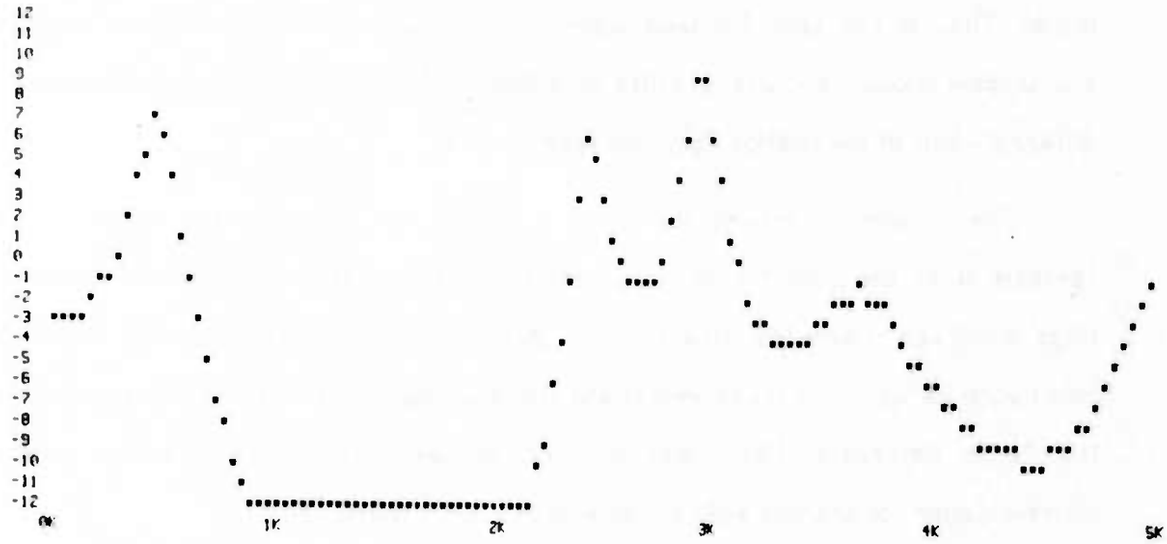


Figure 1.6: SPG Parameters for a 20ms. Window

inputs, that this is only useful when formant bandwidths are also desired.† Fant [Fan74] has indicated that reasonable estimates of the formant amplitudes can be derived simply from their frequencies. Hence, the frequencies themselves seem to be more major information bearing parameters than amplitudes or bandwidths.

1.4.5 Summary

There is a vast range of possible parametric representations, many derived from basic methods of extracting information from the signal. It is not possible to survey the entire field, but we have discussed the methods in common use at the present time. Figure 1.7 summarizes some aspects of the four parametric representations we have chosen.

† Accurate estimates of the formant bandwidths are available from the Covariance method coefficients [Ata71].

Param.	#	Interval size	Window	Approx. bit-rate
SPG	128	10ms.	20ms. Hamming	80K
ACS	15	10ms.	20ms. Hamming	40K
ASA	26	10ms.	10ms. square	15K
ZCC	12	10ms.	10ms. square	10K

Figure 1.7: Four Parametric Representations

1.5 Problem

The main problem to which this research is directed is the comparison and evaluation of parametric representations for speech and their effects upon the performance of speech recognition schemes at the acoustic level. Enough background has been presented now to discuss limitations upon the problem, dimensions of the investigation, and goals of the research. The task implied under the broad statement above is beyond the scope of this dissertation, and is, in view of the lack of clearer models of the entire speech understanding process, beyond the state of the art of performance analysis. Thus, the primary message of this section is how we may limit the analysis so that the results will be meaningful, useful, and extendable to specific system analyses.

1.5.1 Limitations

The first and foremost limitation is to consider only the acoustic level, and at that level, to consider only sources of knowledge that do segmentation and labeling of the input utterance into an acoustic-phonetic transcription. It is reasonable to make these restrictions. The acoustic parameters are primarily input to this level only, although, occasionally, knowledge about such aspects as prosodics will be employed by higher levels. So the main effect of the parametric representation is felt through its effect on the segmentation and labeling processes. Therefore, this effect can be understood to a large degree if the interface between these processes and the rest of the system is understood. That interface is best characterized as a machine transcription.

Second, the acoustic level processes will be measured as a separate sub-system,

with the interface strictly viewed as a transcription of the utterance with boundaries marked in time and some encoding of the identity of each segment. In this way, the interface is more clearly understood and available to analysis. A performance model can be constructed that produces such transcriptions if the two process of segmenting and labeling are able to be modeled individually.

By way of describing the general aspects of the experimental set-up, the following are relevant dimensions from the speech understanding system goals in the Study Group report [New71]:

> 1) Continuous speech is to be used. The articulatory targets, and hence, the resultant acoustic patterns, are much less well achieved in continuous speech than in isolated words. The labeling errors are considerably different therefore, and segmentation becomes harder as well.
>
> 2 -- 5) Cooperative speakers will be used, recording over a high quality microphone in a quiet room. The relaxing of these restrictions may provoke errors, but it is likely that these errors will be predictable in nature -- a general degradation due to many speakers, fricative confusions due to loss of high frequency information, *etc.*
>
> 6 -- 7) Tuning of the acoustics level knowledge will be in the form of pre-testing training data. The training will be over each speaker's utterances, although not the same utterances as used for testing, and thus will be tuned to his voice.
>
> 8) The vocabulary will be chosen to include a wide range of contexts for all the commonly occuring allophones of American English phonemes. [Sho74b]

1.5.2 Performance Dimensions

There are essentially three dimensions to the investigation of the performance of acoustic representations. The first, obviously is the choice of the representation itself. Here, the major task in defining the research is to isolate representative methods from

among the many possibilities.† Although the possible combinations of filter arrays, waveform measurements, spectra, *etc.*, are numerous, the representations that people have chosen thus far seem to fall into a few general types. It is our intention to represent those types according to currently available techniques. If someone invents a new representation for speech, this research will be available to help place the new representation into the total picture. The general types of parameters are: simple measurements on arrays of filters to obtain rough spectral information or to separate different information bearing parts of the signal, LPC parameters of various kinds to parametrize a model of the waveform either acoustically or probabilistically, and spectral envelope estimates that seek to characterize the vocal tract response separately from the excitation source.

Estimating short-time spectra by the output of an array of bandpass filters is represented by the Zero-Crossing count (ZCC) parameters used in Hearsay I [Erm74b] and the Audio Spectrum Analyzer (ASA) [Kri75]. The former consists of five broad bandpass filters with both peak-to-peak amplitudes and zero-crossing counts to increase the ability to estimate frequency information. The latter consists of 25 narrow bandpass filters whose output energy is measured. The LPC method developed by Markel [Mar72] (the autocorrelation method) is used to provide inverse filter coefficients and an estimate of the spectral envelope (SPG) by use of an FFT algorithm. Itakura's log ratio measure [Ita75] will be used in conjunction with the autocoerrelation sequence (ACS), although this representation will not be used with other classification metrics.

The second dimension concerns the particular algorithms used to perform the tasks of labeling and segmentation. These will be based primarily upon the pattern classification concept of a pattern space distance metric. Some traditional metrics employ first or second moment statistical estimates of sample populations of patterns. Two specially

---

† We realize that, no matter which parametrizations are chosen, someone will be sure to point out, "...yes, but if you use *this*, different measurement, you can disambiguate *those* phonetic classes..." The answer to such comments is usually a question, "At what cost; and with what new errors introduced?"

designed metrics will also be used.†

The third dimension can be characterized by the issue of cost. It involves cost of implementation, memory and processor requirements, and the effect of these demands upon total system speed and size (the real-time question). As these issues are much better understood, especially at this level where straightforward, uniform procedures are usually employed, no attempt will be made to span this dimension with empirical results.

1.5.3 Goals

Necessarily, the goals of this research are limited to understanding the effects of parametric representations on acoustic level performance. Central to that understanding are two issues which may be taken as goals.

1) The answers to designer-voiced questions should be available. They are usually of the form, "How much can I get for a certain amount of resource expended?" or "Will I be satisfied (i.e., will the system I am planning be able to use the acoustic level information)?"

2) A methodology for testing and comparing these representations should be available. New representations can, thus, be viewed in perspective. Advances in the state of the art will be recognized and effort can be directed more usefully. This requires a set of algorithms for parametric level processing that are relatively independent of the choice of parametric representation.

1.5.4 Summary

In this section, we have attempted to define a region of the space of possible performance experiments at the lowest level of speech recognition. The entire chapter was aimed at fixing a point of view and a set of basic assumptions about speech understanding systems, the parametric level of analysis, and performance evaluation goals.

---

† Baker's log probability estimate and Itakura's log probability ratio

In that point of view, parametric analysis is the basic input for the lowest level of recognition activity. That activity is primarily performed by segmenting and labeling processes, which produce more manageable data for higher level knowledge sources. When those knowledge sources use knowledge from such high levels as semantics or pragmatics, we may truly call the system a speech understanding system. By carefully evaluating the performance of the low level recognition processes, we may provide a firm base for total system peformance analysis. We have limited this research to a number of the most commonly accepted methods for parametrizing the acoustic signal, and for doing segmentation and labeling. The results and methodology thus provided will further our understanding of many of the issues of speech recognition activity at the parametric level.

Chapter 2

Pattern Classification Techniques

This chapter consists of a short survey of pattern classification and commonly accepted techniques. (For further details see [Dud73, Nag68, Mei72, Fu68].) In chapter 3, we will discuss the ideas from pattern classification theory chosen for this research, the issues surrounding a choice of classes, and considerations for training and testing data corpi.

2.1 Basic Model

Most pattern classification problems are concerned with classifying input patterns into one of a finite number of classes. One approach to pattern classification is to keep a representative of each class, and to match the input for some "closeness" measure with each. This has many shortcomings, not the least being the lack of a way of defining a good template for the various occurrences of speech phenomena under different conditions †. A more general model, for which template matching is a special case, is usually presented. A series of measurements are made on the pattern, either in its original physical form, or from some representation of it. These measurements should be chosen for their invariance under the kinds of informationless perturbations expected and for their dependence upon the classes sought (information content).

Assuming a reasonable set of m features is chosen, their values represent a pattern vector in an m-dimensional feature space. The problem is then to provide a partitioning of that space. (If continuous valued classifications are required, a mapping into the class space is needed.)

A number of different techniques are available for drawing these partitions. Some,

---

† -- The approach has been used for word identification in the Vicens system at a higher level [Vic69]. First the word is segmented and the segments are classified, then the duration-normalized sequence of labels is matched with stored templates for each word in the lexicon.

by nature of their returning a decision value related to an estimate of the confidence or closeness of class identity, can be used to provide continuous classification. Often, however, these values have little meaning outside of their application in partitioning. One usually assumes a single class identity or an ordered subset of the classes (perhaps with estimates of goodness) is to be returned by the classifier.

Various aspects of the acquisition and refinement of these partitionings are of importance. We will discuss the size of sample and test sets of identified patterns and their relevance to the expected results of a method developed with such sets. Algorithms for automatic learning are also available. In these, a teacher is sometimes postulated who can provide feedback to re-adjust the partitioning rules in light of errors committed. Often the set of classes is not known, and unlabeled samples may be partitioned by optimizing various measures of clustering or separability.

By way of example, a simple pattern recognition scheme might work thus:

> Collect, properly segment and label a set of sample patterns (training set)
>
> Average the feature measurements for each class.
>
> For another set of labeled samples (test set) compute the Euclidean distance to each class average from the input features and assign the closest class as the input identity.
>
> If the classification is wrong, adjust the correct class's average towards the new input by 1/n of the distance (where n is the number of training samples in that class). Also adjust the other classes which were closer than the correct one away by a similar fraction.

Obviously, a great many issues are untouched or oversimplified by this example. But it does serve to point out a typical approach. We can easily show that the decision boundaries thus drawn are linear. It has been shown that under certain conditions [Nag66,

Nag68] the adjustment described here converges. With some pre-processing for normalization, this method can provide good results for well clustered classes.

2.2 Stochastic Patterns

Implicit in almost every investigation of Pattern Recognition is the assumption that non-deterministic (stochastic) processes are at work, adding noise and otherwise transforming the original patterns. Let us model this process by asserting that each class corresponds to a multivariate probability distribution in the feature space. If the set of classes corresponds exactly with the information intended to be conveyed by the patterns, this will be a good model. If not, there will be in the observed distributions effects of other sub-class distributions† or of correlation between the classes (in effect, clustering of the clusters)‡. However, we may take this model as a first approximation for speech, although we must investigate the distributions carefully.

For the following development, let:

p_i be the a priori probability of an occurence of a pattern in the ith class.

f_i be the probability density function for the ith class

x be the unknown pattern vector.

then:
$f_i(x)=Pr\{x|class\ i\}$

$p_i*f_i(x)=Pr\{x,class\ i\}$

Bayes rule states that the largest expected rate of correct classification is attained by classifying x in class i if $Pr\{x,i\}\geq Pr\{x,j\}$ for all $j\neq i$. Furthermore, we may define a loss function L(u,v) as the cost of classifying an input in class u when it should be v. Then the expected Loss, or Risk, of a classifying rule C(x) is:

---

† --Multimodality of the cluster for a diphthong, or for vowels in different contexts

‡ --The broad classifications of vowel, nasal, fricative, etc. are much easier to effect than more specific phonetic classes.

$$R = \underset{i}{E}\,\underset{x}{E}\left\{ L[C(x), i] \right\}$$

$$= \sum_{i=1}^{N} \int_{x} L[C(x), i]\; P_r\{i, x\}\, dx \qquad (N \text{ classes})$$

$$= \int_{x} \sum_{i=1}^{N} L[C(x), i]\; P_r\{x, i\}\, dx$$

If we wish to minimize this then we must clearly minimize:

$$\sum_{i=1}^{N} L[C, i]\; p_i\, f_i(x) \qquad \forall x$$

where x is classified in class c.

Until more is known about the relationship between a particular speech understanding system and the classifier it uses, we would assume the first case above which corresponds to a loss of 0 for correct and 1 for incorrect classification. The successful application of Bayes' rule rests upon the availability of the underlying probability distributions. However, they may be estimated parametrically if their forms are known, or approximated by a number of techniques.

2.3 Overview

The methods for estimation of distributions, learning of parameters, and decision boundary drawing may be placed into a few group that will serve to clarify their relationship to the basic model and to optimality as represented above.

If the forms of the distributions are available, we may seek to estimate them parametrically by taking relevant statistics of the samples. For instance, if we have good

reason to believe that the features are independent variables and the clusters have Normal distributions, the variances and means of the features will yield an optimal rule. Normalize by the variances and decide upon the distance to the mean in the normalized space. These are essentially spherical clusters.

Where forms are not known, a number of methods are still available. The method of Potential Functions [Ais64] forms the sum of a number of peak-like functions † each placed at a particular sample point in the class cluster. The amount of spread of each peak determines the smoothness. Many heuristic methods may also be thought of in this light. The kth-nearest neighbor method retains all the samples. The probability is essentially estimated by the number of samples in a class that lie close to the unknown point.

Some decision rules may be thought of as ignoring the distributions and, rather, seeking to find good separating boundaries directly. Forms are chosen, as in the cases of linear or piecewise composite boundaries. Then parameters are estimated from the samples. Equivalences between a number of methods can be shown theoretically.

Learning approaches seek to adjust the parameters of whatever methods are chosen as new information about the patterns is obtained. Supervised learning can occur when a correct label is available for the samples upon which learning takes place. When completely unknown samples are presented, unsupervised learning methods can still obtain rules that separate the samples according to the way they cluster ‡.

A number of transformations upon the pattern space may be made to simplify the task of the recognizer. This is really a continuation of the basic pattern recognition problem, but many researchers have chosen to separate the search for good feature spaces from the search for good decision rules.

---

† -- A spherical Gaussian distribution is often used.

‡ -- Clustering is a concept that must be defined mathematically for such learning to take place.

2.4 Estimating Distributions

Since the Bayes optimum rule assures us the "best" results attainable for a separating boundary decision rule, we would like to be able to apply it. Unfortunately, we may not know the probability densities or the a priori probabilities of the classes. However, if there is some evidence from the nature of the feature measurements, or from the underlying pattern process itself, we may be able to estimate the a priori probabilities and to make assumptions about the form of the densities. This information might also come from statistical analysis of the samples such as estimates of closeness of fit to well-known forms.

The mean vector and covariance matrix fully specify a multivariate normal density function. However, to compute the density values, the covariance matrix must be inverted. The density function is:

$$f(x) = \frac{1}{(2\pi)^{m/2} |C|^{1/2}} \; e^{-\frac{1}{2}\{(x-M)' C^{-1} (x-M)\}}$$

where |C| is the determinant, C the Covariance matrix (mxm), M the mean vector (m), and x the samples (m).

The classes may be composite clusters of a number of forms or they may correspond to highly complex distributions which no simple form can suitably estimate. In fact, we may not fully understand the underlying physical process well enough to derive the form at all.

An important approach available in such a case is that of Potential functions (or Parzen estimators) [Ais64, Mei72]. The estimating density p is directly constructed by superposition of a number of potential functions f as follows:

$$p(x) = \frac{1}{N} \sum_{j=1}^{N} f(x, y_j)$$

Thus, if these estimators are formed on each of the N sample points y, the "density" value of a point x is a superposition of its relation to all the points in that class's sample set. A typical form for f is the multivariate normal with covariance matrix a multiple of the identity matrix (spherical shape, independent dimensions) and mean equal to y. The multiple of identity used for the variance determines the sharpness of the peaks at each point and, thus, the smoothness of the overall function.

Although the Gaussian is very well-behaved, a more computationally efficient function given by Meisel is f(x,y)=h[d(x,y)] where d is a distance function (e.g. the city block function or the sum of the absolute differences in the m dimensions) and h is a piecewise linear window function such as shown below (Figure 2.1).

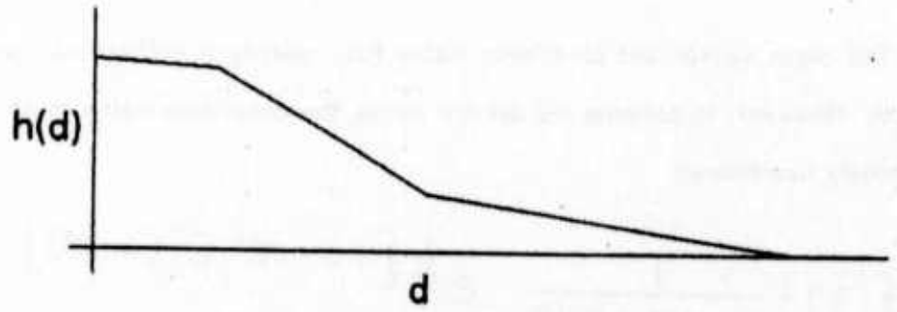


Figure 2.1: A Piecewis-Linear Function

In addition, if $\int_x f(x,y)dx=1$ for all y, then p(x) is guaranteed to be normalized if f is. We can insure that the space is not warped.

Four criteria for the function f are given:

1-f(x,y) should be maximum for x=y

2-f(x,y) should be approximately 0 for x distant from y

3- f(x,y) should be a smooth (continuous) function of distance from x to y

4-if f(x1,y)=f(x2,y) x1 and x2 should be equally similar to y in some sense

One may choose to construct an indirect approximation to the density function. Certainly, a conceptually simple approach is to divide the feature space into small volumes and collect a histogram of the sample patterns. This can be disastrous if the dimensionality is high, however. The number of small volumes in m dimensions is r^m if each feature is broken into r parts and there are m features.

Another problem of the histogram is that it depends very greatly upon the choice of volume chunks for buckets. The sample points may be unimodally distributed, but if the buckets are chosen to split the mode, and if not enough samples are available, spurious modes may be observed. Another technique is to collect the Empirical Cumulative Distribution Function. The N samples are ordered and plotted at increments of 1/N against their values, for the single variate case. The resulting distribution approximation may be smoothed and the slope measured for a density function. The advantage can be seen if one notes that the ECDF depends in some sense on all the points that produce the ordering rather than the points in a single bucket for the shape at any particular location. It is thus a cumulative estimate rather than a local one. Techniques exist for estimating the closeness of fit between standard distributions and an ECDF [Wil68]. Unfortunately, extending the concept to the multivariate case is difficult and not usually done. However, if there is reason to suspect that the features are independent, the ECDFs can be made separately on each dimension.

2.5 Linear Forms

The simplest form for any decision boundary that partitions the feature space into two separate parts is a linear form. Linear discriminant functions have a number of points of appeal and have been extensively investigated†. Often good linear discriminants may be determined as approximations or special cases of more general forms. The advantage of simplicity in the decision rule is apparent to anyone who considers producing a classification of a speech signal into one of 50 classes each 10ms. window, using a vector

† -- Nagy [Nag68] presents a theoretical comparison of a number of linear functions for a two class problem.

of 128 FFT spectral values. Computationally, a linear decision rule can be effected with m multiplies and a comparison, for m dimensions. † The N class problem may take from log2 N to N-1 decision boundaries. Thus, at least 80,000 multiplies per second would be required for the above classification scheme to be done in real time.

An additional advantage is that linear rules are easily parametrized, and thus, learning can take place by means of the adjustment of a reasonably small number of parameters. The effects of transformations of the feature space are more easily understood in connection with simple rules. Finally, higher order forms for decision rules can be reduced to linear form with addition of extra dimensionality.

2.6 Distances

The first thing that comes to mind when considering the pattern space clusters for classification is to somehow use a distance measure from the unknown to the clusters in the decision rule. A number of distance measures have been defined for this purpose. Although we are now faced with slightly more computing, since we must make a distance calculation for each class as opposed to successive dichotomies of the space by boundaries, this approach is more easily adapted to different sets of classes. We need not depend upon the fortuitous placement of clusters where one decision can discard a large set of classes. Furthermore, the usual properties of distance measures ensure that there are no areas of the space where no class identity is assigned.

Euclidean distance is defined, in m-dimensions, as:

$$d(x,y) = \left[\sum_{i=1}^{m} (x_i - y_i)^2 \right]^{1/2}$$

---

† --If the hyperplane dividing the classes has the equation:

$$\sum w_i x_i + \lambda = 0$$

then it can be seen that all vectors x where w.x = -λ lie on the plane. Hence the decision rule is to form the dot products and compare with λ. The distance to the hyperplane is also easily calculated as |w.x+λ|/||w||.

the locus of points such that $d(x,m^1)=d(x,m^2)$ (i.e. the boundary between two classes represented by m^1 and m^2) is:

$$2\sum_{i=1}^{m} x_i (m_i^1 - m_i^2) + \sum_{i=1}^{m} (m_i^2)^2 - \sum_{i=1}^{m} (m_i^1)^2 = 0$$

a linear equation in x. Geometrically, the boundary is a hyperplane which passes perpendicular to the segment joining m^1 and m^2 through its midpoint.

Correlation is defined as:

$$d(x,y) = \frac{x \cdot y}{\|x\| \, \|y\|}$$

or simply the cosine of the angle in m-space between the two vectors. The boundary between two classes here passes through the origin and is at right angles to the plane containing the origin and the two representatives.

More complex forms may yield a distance measure which is still linear. The approximate maximum likelihood method assumes a single covariance matrix for all the classes and multivariate normal distributions. In this case, it can be shown that the locus of equal probability between two classes is a hyperplane cutting the segment connecting the means at the midpoint but not necessarily perpendicular. The equation is:

$$wx - c = 0$$

where: $w = (\mu_1 - \mu_2)' A^{-1}$ $\quad A \equiv$ covariance

$c = \frac{1}{2}(\mu_1 - \mu_2)' A^{-1} (\mu_1 - \mu_2)$ $\quad \mu_j \equiv$ mean

and the distance measure is :

$$d(x,y) = (x-y)' A^{-1} (x-y)$$

There is a more computationally costly procedure than computing from the hyperplane equation, sometimes called the Mahalanobis distance [Dud73] which may be used.

2.7 Piecewise-linear Discriminants

Since linear discriminants are computationally inexpensive, we may wish to define areas of the feature space in which different hyperplanes are used without fear of excessive cost. In fact, arbitrarily complex boundaries can be approximated in this fashion.

The Perceptron model [Ros57] takes just such an approach. Perceptron networks are oriented towards learning linear boundaries between two classes. When applied to the N class situation, the boundaries are all applied in a pairwise fashion, and the classification is made upon the advice of a number of classifiers. Figure 2.2 shows such a situation. Note that the assignment is made when an input pattern gets at least two votes in this case.

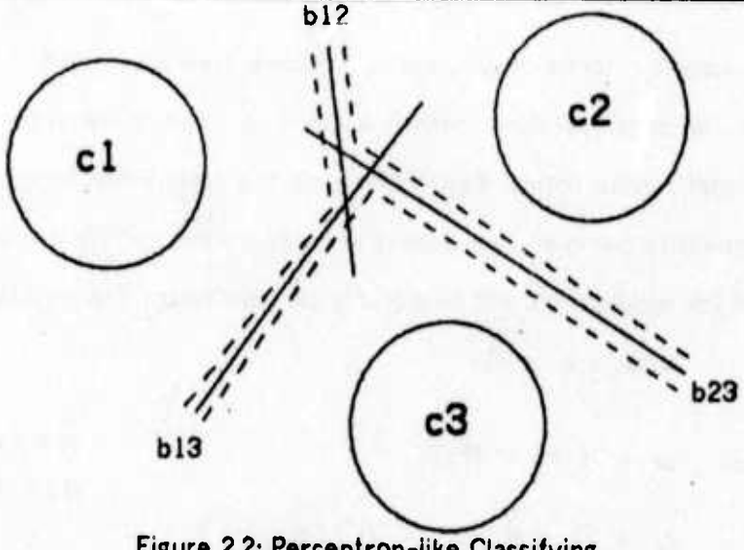


Figure 2.2: Perceptron-like Classifying

We may have a good idea of the implicit subgroups within each class. In this case, we may define boundaries separating samples from each subclass from the rest of the samples in other classes. Then define a piecewise boundary as $Max(w_s.x)$, where w_s are the coefficient vectors of the various subclass discriminants. It can be demonstrated that this forms a convex, piecewise boundary around the composite of the subclasses.

It is clear that certain aspects of linear discriminants are desirable for speech

applications. These include most notably speed, ease of adaptability, and ease of parametrization (size of the learning task). However, some constraints of the speech problem, such as the need to deal with more than two classes, make many of these methods unwieldy to apply or too computationally costly. The amount of computation needed to evaluate the Mahalanobis distance directly is equivalent to that of using the individual covariance matrices in a maximum likelihood rule. Thus, the need to handle many clusters, of unknown placement in the feature space, leads to an algorithmic structure, distance measures, which voids the usefulness of the approximate maximum likelihood assumption. Without a better idea of the actual clustering of the classes, one cannot suggest a generally applicable or even reasonable method for all speech classification, a priori. The conceptualization of the pattern space which most theoretical investigators have brought to the problem may be invalid for speech. For instance, a great deal of overlap may be the inevitable result of variations in speaker, performance, and phonetic context. This is certainly borne out by the difficulty which even experienced phoneticians have in identifying speech sounds in certain contexts or under certain conditions. The best characteristic of linear rules is that they provide an extremely simple structure for learning or tracking, and that they are computationally cheap. The ability to adapt the classifiers under the aegis of higher-level feedback will be much more important to the overall task than careful optimization of performance in a static situation.

2.8 Learning and Tracking of Clusters

The sets of samples from which decision rules are deduced and their parameters extracted are called training sets; the process of acquiring the parameters in a statistically proper manner is dealt with by Bayesian learning methods. In an important sense, the pattern classifier is a learning process -- although of a very simple kind where learning may only occur at set times (training). A number of attempts have been made to employ learning processes more directly (e.g. [Dru73, Nag66, Sel63, Uhr63]), to allow a classifier to seek its own parameters or its own rules, to allow it to adapt them to slowly varying pattern spaces, and to allow the identification of patterns for which no classification was expected.

Many algorithms have been set forward which guarantee convergence to an optimal linear classifier under various conditions. A famous one, for which an upper bound on the number of steps for convergence was derived, is the perceptron error correcting algorithm [Ros57]. The training set is presented in order as often as necessary and the coefficients of the linear boundary separating the two classes, C1 and C2, change as follows, for j=1,...:

$w(j+1) = w(j) + x(j)$ if $x(j)$ in C1 and $w(j) \cdot x \leq \lambda$

$w(j+1) = w(j) - x(j)$ if $x(j)$ in C2 and $w(j) \cdot x \geq \lambda$

$w(j+1) = w(j)$ otherwise

where x is classified as C1 if $w \cdot x \geq \lambda$, etc. The weights are thus adjusted on misclassification. A number of variations on this scheme deal with various amounts for adjusting the w vector, varying the order of presentation, and the problem of using imperfect components. Nagy [Nag68] discusses some of these results and also points out that if the problem (i.e. the samples from the two classes) is not indeed linearly separable, learning methods may not converge. Instead, they may oscillate or converge on a local optimum. However, one way of gaining insight into whether two classes are linearly separable is to try such a learning scheme on them.

Related to these methods are devices for tracking varying pattern clusters by adjusting a linear discriminant to follow new patterns as they are presented. It has been assumed so far that the training samples are drawn from the same distribution as the unknown patterns will be. However, there may be overall changes in the pattern producer that shift the clusters very slowly (relative to the frequency of incoming patterns) as one proceeds with actual recognition. Consider the effects of excitement, fatigue, or confidence upon a speaker dealing with a computer speech recognition system. Some speaker normalization may be treated in this fashion as well. Although in the case of changing speakers, the clusters shift suddenly, such a change happens rarely in the time scale we are considering and much structure of the pattern space is common among different speakers. Such adaptive behavior in pattern recognition dates back to work on Morse code [GolB59, Sel63].

While many of the methods discussed here are on shaky footing without an idea of some of the aspects of the particular patterns in question -- how well they cluster, what shape the clusters are, or how many -- there do seem to be areas of applicability in speech recognition. In particular, it may be possible from phonetic studies to predict how many clusters will be found and develop a good idea of the hierarchy of classes into which they may be placed. However, we must take care not to infer relationships upon the pattern space that we feel exist in some linguistic model of speech perception. The best application for these methods is in discovering how well our ideas of the proper set of classes coincide with particular feature spaces, the data, and rules we have chosen. In addition, tracking techniques may well prove useful in going from one speaker to another or in changing acoustic environments. The ability to train without tedious hand labeling of speech data would be a great help, but the assumptions necessary for unsupervised learning algorithms thus far discovered do not seem to apply well to speech.

2.9 Conclusions and Discussion

This survey has taken a narrower view of pattern classification than is sometimes set forth by those who have viewed entire pattern recognition systems (e.g. [Min63, New71, Uhr73]). The gross description of any pattern recognition problem is simply: map a space of patterns onto a space of symbols (usually a finite set of names). The speech recognition problem can be variously viewed as any of the following mappings:

utterances	->	semantic states
phrases	->	syntactic structures
words	->	lexical entries
segments	->	phones
time slices	->	acoustic-phonetic labels

This many level view is taken in some currently operative systems such as Dixon [Dix75a] and Reddy [Red73]. Each level has its own data structures and sources of knowledge from which rules of varying complexity can be deduced -- either for recognition or interlevel translation. The same thing occurs in visual pattern recognition with the various levels proposed for scene analysis and picture decomposition. Speech in

particular, has large bodies of research relevant to these levels, although most of the rules are available in generative form [New71]. Thus the question of where to draw the line between "raw" pattern classification and various processes for reduction of search, inference of goals, syntactic analysis, and even phonetic segmentation is a system structure question. A good decomposition of the patterns at each level exists in the rules that translate to the next lower level. The decomposition into phone-like labels for short time windows is the most primitive one proposed. There is general agreement that the burden of complex processing, feedback, context, etc. should be placed on the higher levels with the stream of labels serving as input to them. This sort of constraint will be necessary in order to achieve real-time response in the future. Furthermore, a translation to symbolic form must be made on the input signal for space and time economies in the higher level processes as well.

The previous survey provides an overview of the methods available for classification. However, some specific limitations must be made, justified by what is known about speech, in order that the comparisons that are the object of this research may be made in a reasonable time frame. The central aim of this discussion is to fix upon a useful environment for comparing parametric representations. At the level of labelling the acoustic signal, that means finding a "typical" algorithm or family of algorithms for classification. The constraints placed upon this choice are time and space requirements, the need for graceful error recovery and robustness under variations in the input quality, and the necessity of experimenting with methods at the level of current technology.

Speech understanding systems have not generally employed extremely powerful pattern classification schemes. The view has been often expressed that the problem must be addressed at multiple levels by a variety of knowledge sources. The benefit of a highly tuned and powerful classifier is nullified by the fact that speech has such inherent variability that even different utterances by the same speaker will demonstrate widely differing patterns for the same informational elements. The warning against overkill in tuning to any set of training samples, no matter how extensive, has been made throughout the pattern recognition literature, and speech is clearly subject to this problem.

We have chosen, therefore, a few, well-accepted, classifiers, simple and robust, that cover much of the current speech understanding usage. The existence of more complex methods at the state of the art should not invalidate these results, since the comparative performance of parametric representations in the context of these straightforward methods will serve as guides to the design of simpler, soon realizable systems for limited tasks, as well as first-order predictions of more complex classifiers built upon the basic classification algorithms. The latter will be true for the following reasons: The complex methods for classifying sounds thus far proposed have been based, perhaps in a hierarchy of decisions, upon simple concepts of "closeness" or "matching" in a pattern space or over certain elements of evidence. Second, the evidence provided by human production and recognition of speech seems to imply a continuum of sounds, arbitrarily interpreted as belonging to information bearing classes. A classifier and parametric representation which adequately capture the structure of that continuum can be expected to perform well in partitioning it. Thus, a parametric representation which allows some uniform distance function to separate patterns according to their actual information class, will have indicated this similar structure, as well as an amenability to the use of concepts of closeness in more complex methods.

The concept of distance is central to the use of a pattern space as a representation of the actual phenomena. Thus the simplest method of classifying a pattern is according to the distance from that pattern to the clusters of sample patterns. If the clusters are parametrized by some M_i for samples belonging to class C_i, then the distance is some function of the input, x, and M_i, and x is classified in C_j, where j minimizes the distance. The property that the distance is a function of a single class's parameters means that no region of the pattern space is unclassified, since every point must produce some minimal value. If there are m classes then m evaluations must be made and m sets of cluster parameters must be stored. A hierarchy of decisions would provide a clear computational improvement, but may be derived after the representation and distance function are chosen, or may be deduced from the pairwise distances of the clusters themselves. In the context of this research, the independent evaluations provide flexibility in the choice of a

set of classes and a way of comparing entire parametric spaces to one another upon a simple base.

Finally, the value returned as a distance may be used to estimate the probability of that class being the proper choice. Some distances estimate the distributions of patterns for each class and provide this probability directly. Others may be compared to empirical distributions of distances for that class. When such probabilities are available, meaningful combinations of classifications can be made and the information available from classes that are close to minimal distance can be employed.

The choice of a limited kind of classifier will not hurt the usefulness of the results because most classifiers are based upon the distance concept. But more importantly, the choice will help because results will be easily applied to simple systems or parts of more complex ones. The role of parametric representations can be more clearly seen in the light of results that measure performance of simple algorithms. And the structural similarity of the pattern space to the space of speech sounds is brought clearly into focus.

Chapter 3

Experimental Considerations

This chapter continues the discussion of pattern classification issues. Two very important issues are: 1) the set of recognition *targets* -- the information-bearing classes, as distinct from the *templates* for those targets -- which are the object of classification activity; and 2) aspects of data quality and quantity. The methodology of training and the experiments we have devised in this research are strongly affected by these considerations.

3.1 Acoustic-Phonetic Classes

A very important issue is the number and nature of the classes that are the output of any classifier. Sometimes this choice is trivial -- in fault recognition in machinery for example, there are two classes, faulty and fault-free. Often, even in binary choices like this however, there are many sub-classifications -- unnecessary in all but a few special cases (such as "fault-free but having a slight vibration that might lead to faults after extended use" to continue the example.) The situation in speech is just such a one. Nasalized vowels, devoiced glides, transition portions, deleted or altered consonants all represent subclasses of things one may wish to deal with most of the time as entire classes. Consequently, we feel some discussion of the directions available is warranted. Our view of speech understanding systems is that non-local considerations, relating to context, speaker idiosyncracies, reduction of search, and knowledge about other levels than acoustic-phonetics, will be dealt with at those other levels by other processes. Hence, the kinds of recognition that some systems have done or propose to do -- tracking slopes of formants, for example to disambiguate consonants by the transitions into following sonorants -- will require different objects of recognition than the process that notices the appropriate context in which to invoke such specialized rules.

Data reduction, representation transduction, and hypothesis generation are the

principal roles for the processes we are investigating. These require general utility over a broad range of speech sounds, robustness, and low cost at least for some (initial) parametric analysis routines.

There are two alternative approaches that have been taken in defining or discovering classes for speech pattern classification, which may be called *acoustic gestures* and *features*. The acoustic gesture approach takes the view that the phonetic significance of a speech segment is to be extracted by other sources of knowledge. The classes into which the segment is mapped have phonetic correlates, to be sure, but those correlates are subject to context and speaker variations as well as the acoustic nature of the segment. A set of gestures is chosen, therefore, which represents all the significantly different sounds encountered in speech. Where the difference between two phonetic classes is clearly reflected in a difference in their acoustic realizations, the task of differentiating them is accomplished by differentiating the corresponding acoustic gestures. Where the same sounds may realize different phonetic classes, however, an optimization of sorts must be made, balancing usefulness of the information supplied by identifying the acoustic gesture with the complexity of a growing number of specialized cases, each identified as a separate gesture and many overlapping in the pattern space. The boundaries, and thus the classes, most suitable to the problem divide the pattern space into regions such that any pattern in a particular region could be a realization of the phonetic situation that produced any other pattern in that region.

The feature approach is well described by Meisel: "The selection of a set of features which efficiently describe a system in terms of a pattern to be recognized in those features is itself a pattern recognition (problem). Each feature describes some aspect of the pattern and amounts to a decomposition of the quality to be recognized in the overall problem into a set of more easily recognized qualities." [Mei72] Since the object is to disambiguate phonetic situations, the features must be those qualities that distinguish different such situations. While the parametric representations may be viewed as features, we would prefer to reserve the term for those qualities that more highly

correlate with phonetic information. For example, energy in the frequency band from 3KHz to 5KHz may often imply the feature "Fricative," yet there are enough cases of non-fricatives, high vowels for instance, which produce moderate amounts of energy in that band. Obviously, a parametric representation which effectively allows for extraction of these teleological features is a good one, but it has been the experience in speech research that such a representation is very hard to find.

Under the influence of the simple model of a classifier used in this research, these two points of view merge. More complex recognizers often mix the two views in hierarchical schemes where the presence of certain features may trigger attempts to classify among a sub-set of the target phonetic labels [Erm74b]. If the gestures correspond to a set of features, then the fact of recognizing a gesture provides evidence in favor of its features' presence. Likewise, the set of features recognized forms an address of the gesture. The duality of this relationship depends upon having the corresponding weights available to calculate how much evidence is provided. (See figure 3.1 for the features and weights used in Hearsay II, phonetic hypothesizing.)

;FEATS:	HI	MID	LO	FRNT	CENT	BK	RND	NTR	VEL	NAS	MIN	VCD	NUL	LOF	FRC	VOC	CON	DIPH
;WGTS:	75	75	75	75	75	75	100	100	100	100	50	100	100	100	100	80	80	100
-	+5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	+5	-5	-5	-5	+5	-5
B	+5	-5	-5	+5	-2	-5	-2	-5	-5	-5	-4	+5	-2	-1	-5	-5	+5	-5
P	+3	-3	-5	+5	-2	-5	-2	-5	-5	-5	+1	-5	-5	-5	+5	-5	+5	-5
F	+3	-3	-5	+5	-5	-5	-2	-5	-5	-5	-2	-5	-2	-5	+4	-5	+5	-5
Z	+3	-3	+5	-2	+5	-3	-5	-5	-5	-5	+2	+5	-5	-5	+5	-5	+5	-5
S	+3	-3	-5	-2	+5	-2	-5	-5	-5	-5	+3	-5	-5	-5	+5	-5	+5	-5
R	+1	-3	-5	-2	+5	+0	+0	+5	-1	-5	+3	+5	-5	-2	-5	-2	+4	-5
AX	-2	+0	-2	-2	+5	+0	-5	-5	-2	-5	+3	+3	-5	-5	-5	+5	-5	-5
AA	-5	-5	+0	-5	-5	+4	-5	-5	+0	-5	+5	+5	-5	+4	-5	+5	-5	-5

Figure 3.1: Some Phonetic Features and Weights, HSII

In recognizing continuous speech, segments are usually labeled with some phonetic information. As more phonetic information is available for a segment and its context, one can state with more certainty the phonemic interpretation of that portion of the signal. An

interpretation is being made and the speech understanding system is making it. Thus, the most important criterion for the source of knowledge which provides the phonetic information is whether it is adequately serving the needs of the rest of the system. Those needs are for interpretations to be made consistently-- certain phones in certain contexts should be interpreted in the same way -- and robustness -- unusual contexts or poor conditions should bring about interpretations which are wrong in proportion to the degree of degradation encountered.

In conflict with recognizing a single phonetic situation is the problem that features interact in complex ways under various conditions. The feature "high third formant" may mean entirely different things depending upon the location of the second formant, or may be irrelevant. These interactions could be expressed as rules and the appropriate recognitions could be the output of a complex interpretive stage. The question is whether anything is gained. The key to a suitable feature decomposition of the pattern recognition problem is the discovery of a set of features which are easier or more accurate (or both). While many methods of speech analysis have attempted to provide parameters that make such recognitions easy, none has been able to carry the entire burden. Some features are easily extracted from one parametric representation and others are just confused, yet all are necessary if critical errors are to be avoided.

If one attempts to interpret the speech signal by the results of acoustic gesture recognition, one finds that different phones often give rise to the same acoustic gesture, and, conversely, that the same phone can be influenced by phonetic context and conditions of emotion, speaker accent, prosodics, and so on, to produce different acoustic gestures. In the same way that features might be disambiguated by a system of rules, so could a string of acoustic gestures by corrected. The context dependent errors will appear in sequence with labels that indicate the context. Alternative labels of acoustic gestures would give an indication of likely candidates for error correction. The sequence /g I e - s/ for example, could be altered to read /g e t/ by application of some simple and obvious rules from the phonetic source of knowledge. In fact, features could be extracted from

the sequence of acoustic gestures. The recognition of /m/ means "voicing","low second formant", and "nasal" features are present to some extent. Again the major issue is whether the recognition process is easier or more accurate.

Taking the entire label as a unit will tend to help avoid errors resulting from interactions with one or two wrong features. The classifier can rank the labels, and thus provide graceful error correction. On the other hand, in order to represent a wide range of acoustic situations encountered in continuous speech, and to provide sufficient disambiguation of similar phones under most conditions, the set of acoustic gestures must be fairly large -- multiple labels for each segment may be required. The trade-offs depend upon system organization to some extent. The amount of higher level support is a significant factor in this choice. However, the essential interchangeability of the two views should be apparent. For the purposes of this research, it does not matter whether the parametric representations are compared for their feature recognition support or their acoustic gesture recognition. Rather, their ability to allow the classifier to come up with the label or feature set best corresponding to the phonetic situation can be measured over the entire test data by either method of recognition.

The labeling results that will be reported are accuracies over a set of acoustic gestures with phonetic interpretations. The selection of these is important, as is the training of the classifiers used.

In chapter 7, we will discuss an algorithm for discovering the natural clustering of a set of sample patterns, and for discovering representativesi -- *templates* -- for each cluster from the samples themselves. This algorithm is used to refine a set of phonetic labels -- *targets* -- which have been marked over a training corpus of speech parameters by hand segmentation and labeling. There are a great many other methods for refining such a set of classes. Iterative adjustment techniques are discussed in the pattern classification literature [Nag68]. In addition, learning techniques might be used to adjust the set as well as to "track" slow shifts in the nature of the data in an operating speech understanding system. A careful survey and analysis of the relative merits of these

methods seems to be rather far removed from the central issues of this dissertation. We are primarily concerned with acquisition of a reasonable set of classes for testing labeling proficiency in a benchmark procedure. The clustering method implemented has given evidence of quite adequate performance, as well as of being consistent with the viewpoint of acoustics-oriented, parameter independence.

3.2 Data Quality and Quantity

The discussions above have been concerned with empirical methods for pattern classification for speech understanding systems. Consequently the data upon which the methods are based is an important factor in the validity of the results. While the form of a decision rule may be chosen by intuition, necessity, or fiat, the data which provide training statistics (which most rules use) is never perfect. The data which provides the testing results must be subject to similar scrutiny. Are the corpi representative of speech (and what does that mean)? Are they large enough? What should be the relationship between training and testing data?

3.2.1 Speech Data

The quality of data one acquires in a body of speech depends upon who is speaking, how he is speaking, and how the recording is made. Many applications and a number of existing systems deal with isolated words, and thus avoid the variability introduced in continuous speech by coarticulation, varying stress, and the difficulty of isolating the proper segments to form a lexeme. Variation in speaker falls into three types -- different speakers have different gross vocal characteristics, the most notable being fundamental pitch frequency; individual speakers within one gross type use different portions of the acoustic domain for particular phonetic items (varying "dialect"); and speakers vary in the coarticulation rules that may explain their performance. Effects of ambient noise, microphone characteristics, and such must be dealt with as well at some level.

Speech recognition research has been primarily concerned with intra-speaker variations. This is not to say that methods for normalizing inter-speaker variations are not

a significant part of the problem, but it is generally felt that a model capable of dealing with one speaker well, will be extendible to the multi-speaker problem. This view arises from the observation that nearly every sound within the usual domain of speech may be expected to be produced by a speaker in some context. Furthermore, we understand total strangers, albeit those with unusual accents cause more difficulty, with very little training effort. Certainly, the acoustic recognition process in human (or machine) perception of speech is little affected by changes of a phonological nature. These are handled (or should be) by higher level transformations. Whereas changes in the interpretation of acoustic gestures from one speaker to another, while seemingly a more fundamental normalization problem, appear to be the kind of problem more amenable to solution by some fairly straightforward iterative adjustment or learning techniques. There must be a common structure to the patterns we process that is very pervasive. The ability of a parametric representation to measure just that structure will become apparent by experiments with no consideration given to speaker normalization. It can be observed that almost as much variation exists within one speaker's performance as between two similar speakers'.

A final consideration is that of recording quality and related issues. Some difficulty is introduced when microphone characteristics, for example, impose varying spectral characteristics, or when noise influences the values of some parameters. However, noise subtraction and spectral leveling methods can be employed, and good quality equipment and quiet recording conditions are reasonable for these experiments.

The following decisions seem reasonable for training and testing data: Continuous speech by adult male speakers will be used for testing. The sentences will be drawn from a variety of task domains and therefore represent various words and sentences for varying phonetic contexts. All recordings will be on reasonably high quality equipment with low ambient noise, but no excessive concern with this seems warranted.

3.2.2 Quantities

It is very difficult to acquire large numbers of sample speech patterns which are properly labeled. This is a common problem to other pattern classification applications, although for a variety of reasons. In speech, the difficulty is in properly segmenting and labeling by hand the training and test corpi. Although refinement methods will help in training with poorly labeled data, the best possible hand segmentation and labeling should be used. Testing can produce valid estimates of accuracy only if the test data is also properly marked, and the validity of the estimates will increase with the quantity of data. There is a strong tendency, therefore to try for as much data as possible, and to make what data is available serve both purposes.

One can over-design or over-train a rule to the point of degrading performance on data other than the training set. This phenomenon depends upon the type of classifier rule, but also upon the fact that the data may not be truly representative. As more data is used, this latter becomes less likely. The size of the training set, then, should relate to the complexity of the classifier. More complex classifiers will be better able to separate the training sample clusters, and thus take on their particular structure, so one would want that structure to be more representative of speech in the pattern space. For example, a nearest-neighbor rule, trained on a few samples, may allow a few spurious samples to capture large areas of what should be another class because no representatives of that other class were in the training data. A simpler, Euclidean distance classifier using the sample means, would be less affected by the few bad samples.

The number of dimensions of the pattern space also plays a role.. Data points in high dimensional spaces will be remarkably far apart and thinly spread. If one wishes to estimate a distribution in such a space, one needs a large number of points to fill in a histogram, fewer to produce a valid covariance matrix, and fewer still to estimate the mean. The simple classification rules may be as good when trained on a few samples as on many for these reasons.

Meisel [Mei72] points out that the ratio of the number of samples, N, to the number of features, m, should be significantly larger than unity. His experimental results indicated an N/m of 3 - 5 usually yields successful training. However, consider the case where 100 two-valued patterns are available. If someone maliciously adds 98 values to each pattern, all drawn from the same distribution, the results will still be as good as in the case where N/m was 50. Thus the "true" dimension of the patterns must be considered. Meisel also gives as example an experiment where 10 two-valued samples are drawn from the same distribution for each of two supposedly different classes. A linear transformation could be found which allowed a perfect classification with a linear boundary. Yet, when 90 more samples were produced, no such transformation could be found. The two clusters had merged perfectly.

The expected performance of rules in practice can only be estimated by examining their performance on separate test sets. (If distributions are available for the expected inputs, theoretical bounds can be derived for the expected error rates.) Either the available samples must be divided into training and test sets for this purpose or else an iterative process as follows can be used.

Divide the samples into k sets: $S_1 \ldots S_k$

Train on all but S_i and test on S_i, for each i.

It can be argued that if k is the number of samples, this is equivalent to testing on the training set. Yet there are some methods -- Potential functions or Nearest Neighbor -- which are guaranteed to perform perfectly on the training set that will not necessarily do so in this case.

The training sets to be used in this research consist of enough data to provide between 30 and 150 sample segments for each of about 40 common phones of English. While this is not really enough to provide good estimates of all the major allophonic variants of each phone, the difficulties of acquiring the extremely fine hand segmentation and labeling needed to avoid introducing errors in training have precluded using more data. While this gives a rather small N/m value for m=128 (in the SPG case), it must be

noted that the SPG parameters are highly correlated with one another and are, in fact, derived from 15 LPC parameters. No doubt the "true" dimensionality of the space is considerably lower.

The test data is representative of the kind of speech to be encountered by actual speech understanding systems. For the near future, that seems to be cooperative speakers, high quality low noise conditions, somewhat limited vocabularies, and continuous speech. To that end, we have chosen to train with a separate set of 27 sentences spoken by each speaker. These contain approximately 1200 phonetic segments, and are designed to contain a number of instances of the most commonly occuring allophones of English. [Sho74b]

3.3 Summary

We have discussed two issues of considerable importance to recognition of speech: recognition targets and data quality and quantity. We have tried to make choices in these areas which are reasonable, given the limited resources available. Considerably more data is being dealt with in this research than has been the case in past efforts. Training and testing sets are of the order of 1000 segments and are separate data. Recording is still over high quality microphones. The choice of template recognition instead of feature recognition is one of avialable methods, and personal preference.

Chapter 4

Speech Recognition Systems

It is impossible to study the parametric level without paying some attention to the total systems. In recent years, there have been a number of implementations of speech recognition systems with a variety of knowledge sources, control mechanisms, and data structures. Some of these systems have understanding of the content of the utterance as part of their power. Others may be called word recognition systems and are oriented towards isolated word recognition by uniform strategies, or with limited knowledge sources. All these systems have a component which analyzes the parametric input. Almost all produce a phonetic-like transcription of the utterence as some internal intermediate representation. This chapter is a brief survey of the more sallent aspects and the available performance measurements of a number of systems in current developement.

4.1 The Parametric Level

The previous discussions have dealt with the individual aspects of acoustic-phonetic processing in speech understanding systems: the parametric representation, the roles of segmentation and of labeling, pattern classification techniques, costs, and quality and quantity of data. In one sense, this covers most of the background material for the particular experimental results to be presented. However, it is often important to view such results with the perspective of other research in the area. Indeed, a large part of the effort spent in this work was spent in developing methods which could perform reasonably well by current standards and yet which would not be specific to any particular parametric representation. This chapter is a brief survey of the parametric level analysis performed in a number of current speech understanding systems or partial systems. However, it is almost impossible to meaningfully compare their performance results in a quantitative manner since they use different representations for output of the

acoustic-phonetic information, since the published results involve widely varying test conditions and methods of evaluation, and since the design goals of the various programs differ considerably depending upon the structure and goals of the complete systems of which they are sub-parts. This survey merely seeks to provide a framework for better understanding of the role of acoustic parameters, their use, and the kind of performance currently available. In addition to the mechanical speech recognition work, we will cite some human performance results [Sho75a] in order to better place machine performance in perspective.

Since we have had only limited contact with many of the systems currently being developed, we have had to rely upon published descriptions of their methods, performance, and overall structure. In this, we were greatly aided by an in-depth survey of four acoustic-phonetic levels of large systems by Hieronymus [Hie75]. However, some assertions and methods of evaluation which have been encountered in the speech recognition literature seem to be biased by pre-conceived notions, or to reflect poor techniques. For example, Hieronymus, in summarizing vowel recognition for the system at Lincoln Labs, points out that 50% vowel identification accuracy for first-choice is very poor since humans find this task so easy. Yet Shockey and Reddy discovered that human vowel perception -- using auditory input -- of continuous, but unfamiliar, speech was not significantly better, perhaps 55% to 60%. This mistaken belief that humans do very well because they have some spectacularly successful auditory mechanism has doubtless led to a great deal of misdirected effort.

A number of published results appear to have been based upon testing with the same data corpus used for training (if training is done) or gathering of statistics to describe recognition targets. Even "tuning" or deriving rules from the same data used for testing can bias results. It is an understandable mistake to make since much speech data seems to us to be of equivalent difficulty and quality. Yet this is a bad practice. We have observed considerable degrading of performance results when separate test data is used, indicating that the results over the training data are artificially high. This is a point that is

strongly taken in the more traditional pattern classification literature. The manner and extent of development of classifiers can severly bias results.

A final point on the difficulty of directly comparing reported results is that the application of phonological knowledge greatly improves some of the raw acoustic recognition results before they are measured. This is unavoidable since some systems have phonological mechanisms built in at the lowest level, while others apply this knowledge at other levels -- either explicitly or implicitly (e.g., integrated into the lexical entries) -- or not at all. One experiment with the Dragon system has shown us just how much action can result, in continuous speech, from this knowledge. When a set of templates for phone recognition, acquired by the clustering algorithm to be described in chapter 7, was used, word accuracy of the system dropped dramatically from previous results. Yet when new phonological descriptions (all drawn from inspection of the training data only) were integrated into the word lexicon, performance was improved to better than 95% over a separate test set of sentences.†

There are, therefore, a variety of reasons for the relative lack of comparative studies at this (or other) levels of speech understanding systems. The least that a look at the current efforts can show is the essentially central nature of the parametric representation, the important role often played by pattern classification techniques, and a fairly broad consensus on the role of acoustic-phonetic analysis in the total understanding task.

4.2 Large Systems

The following descriptions are, of necessity, very brief. We cannot hope to do justice to the great deal of effort and knowledge that has gone into these programs. We are merely trying to gain a general understanding of the nature and quality of recognition at the lowest level of a number of current speech understanding systems. In a number of

† Clearly the prior lexical entries contained valid phonological descriptions of the words in terms of the old templates, but when a new set of templates was introduced, a new phonetics was introduced, hence the need for a new phonology.

cases, labeling accuracies are cited for classes of sounds such as vowels or fricatives. We do not know how many phones are contained in each of these classes, so we cannot recommend detailed cross-comparisons of these results.

4.2.1 BBN Speechlis

The Bolt Beranek and Newman Speechlis system [Woo74, Woo75, SchR75] acoustic-phonetic recognition procedures are based upon a lattice data representation used throughout that system, which allows alternative segmentation and labeling decisions to be maintained. They have chosen a number of fairly feature-specific parameters (i.e., designed to capture specific phonetic features) which are input to a set of heuristic decision procedures. The stated philosophy is to deal with the inherent ambiguities in the speech signal by allowing ambiguity in the recognition process. First, segment boundaries are located by looking for clues in any one of a set of special segmentation parameters, then labeling is performed on a different set†, averaged over the central half of each segment. This produces a broad class label from [Sonorant, Obstruent, Fricative, Nasal, Plosive]. Finally, class-specific decision procedures are applied to identify each segment as one of a set of 36 phones.

Hieronymus reports segmentation accuracy of about 5% missing and 5% extra boundaries on a small set of data. Labeling accuracy is about 91% correct vowel identification within three choices.‡ Formant tracking and speaker normalization functions are employed to benefit here. It is not known whether this accuracy figure includes cases where the vowel identification routine was not invoked because of an incorrect class label. This does appear the most sophisticated aspect of their labeling.

In general, the structure of the program seems interesting, the invocation of special tests, formant normalization, and the alternative segment structure being often cited as

---

† LPC parameters such as formant frequencies seem to play an important role in both processes

‡ Vowel identification (presumably after the vowel segment has been located and classed as such) is often cited as an important statistic. It sometimes is the only labeling process that can be separated from segmentation or phonological analysis.

powerful techniques to be applied at this level. However, we suspect that these preliminary results should be treated carefully until more detailed performance analysis is forthcoming.

4.2.2 CMU Hearsay I and II

There are three systems being developed at Carnegie-Mellon University: Hearsay I, which is not being carried any further, Hearsay II, the research system for a great deal of the current effort, and Dragon and related systems, developed by Baker [BakJK75b] and extended by Lowerre [Low76]. We will discuss Dragon later in this chapter.

Hearsay I [Red73] was developed to test the Hypothesize and Verify paradigm as well as to provide a system which give relatively great importance to higher level knowledge sources. The acoustic parameters are six derived parameters based upon the amplitude and zero-crossing measures from octave bandpass analog filters (ZCC mentioned earlier). A pseudo-phone (PP) label is placed on the signal every 10 ms. The stream of PPs is smoothed and their broad class memberships† yield a first segmentation. Then some correction and further segment identification is made using additional parameters designed to measure overall energy and locate points of maximal energy such as vowels. The same classification function is used to verify phones hypothesized by higher level knowledge sources as is used to label the segments. Labeling is based upon the Euclidean distance from the input 10 ms. sample parameters to a set of speaker-specific templates for the PPs. These are trained on a list of neutral-context words, one training segment per label.

The acoustic-phonetic processing in Hearsay I is fairly crude, and the good performance of the system is, to a large extent due to correct application of syntax and semantics.

The procedures developed for use in this thesis research are being employed as the parametric level for Hearsay II [Erm75]. Some of the parametric representations which we

† Silence, Fricative, Voiced

test here are currently being considered. This is not a critical design decision, however, because of the flexible nature of these routines. The set of phone-like labels for the acoustic-phonetic transcription output of this level are re-processed at a higher level. A set of ternary phonetic features is assigned to each label and weighted according to their relative importance. A simple algebra of features and weights is then available to combine alternative labels with scores, and to convert back to phonetic labels if so desired. This information, along with stress contour analysis to locate syllabic nuclei, location of stop consonant patterns in the sequence of acoustic segments, and recombination of similar features, produces hypotheses of segments which may overlap in time or have multiple alternative labels (much as in the BBN lattice).

A major concern is to avoid disastrous errors by paying the higher cost of keeping many hypotheses around. Thus, the segmentation is tuned to miss as few boundaries as possible (approximately 2%). The use of phonetic features at the next level, where recombining of segments as well as hypothesizing and verifying of labels is done, allows the partial match of correct phones where a less conservative system might reject them. A final note about Hearsay II is that it is probably the most flexible overall system organization being developed for speech understanding. There is nothing in the global data representations nor in the control structures to preclude applying knowledge at any time to various parts of the data. Such decisions are made by the knowledge sources themselves in an asynchronous fashion.

4.2.3 SDC VDM System

The VDM System of System Development Corporation is oriented toward verifying, at the acoustic-phonetic level, a string of phones hypothesized by other levels. [Rit74] Lexical entries are used to generate phonemic, then phonetic, and finally parametric representations of hypothesized syllables through the application of lexical lookup, phonological rules, and parametric mapping procedures. Then a syllable mapping process attempts to match the observed parameters with the hypothesized ones. (Actually, word beginnings are found in a bottom-up segmentation.) A number of coarticulation rules are implemented in a very flexible structure.

Vowel recognition is reported at 77% for three choices (48% for first choice) with an additional 11% very nearly the correct vowel. Segmentation of vowels, extremely important in a syllable-oriented system, was 91% correctly found. Fricatives (except /th/) were generally in the 80% recognition range.†

As an overall view, there seems to be a great deal of phonological and coarticulative knowledge being applied by this system as an integrated part of the parametric recognition processes. This makes it difficult to use the reported results in direct comparison with other parametric level recognizers. Some experimental results for individual pieces of this system are available [Mol74, Gil74, Kam74] but the conditions of testing vary in quality.

4.2.4 MIT Lincoln Labs

The speech understanding system developed at MIT Lincoln Laboratories seems, at this time, to have the most sophisticated acoustic-phonetic analysis of the systems mentioned thus far. [For74, Wie74] It is a bottom-up system with phonetic segments and labels being produced from the acoustic parameters without help from higher level knowledge sources. Essentially, segmentation is performed and formant tracking proceeds first. Pitch and frication detection is also done. Then a broad assignment of [Vowel-like, Dip, Fricative, Stop] is made, and specialized identification routines are applied. In the case of Vowel-like segments, some further segmentation of semivowels and other voiced consonants is performed. A final stage merges and edits the various decisions made thus far.

The results of labeling vowels are 41% and 69% for first choice and first three choices, respectively. Dips were recognized with 82%, and fricatives with 91% accuracy. It must be noted that these classes were rather broad. A single class each represented nasals, glides, and liquids. There was however, careful attention given to testing conditions, the data was described in the reports, and one is inclined to believe that the above results are reliable estimates of the performance.

---

† As more details of the testing were reported, these results may be more reliable than some others.

4.2.5 IBM Research -- GLODIS with Speech Knowledge

The IBM Speech Processing Group, using the General Language-Oriented Decision Implementation System (GLODIS), has implemented a number of heuristic rules for phoneme identification. [Dix75a] The inputs to this system are a digital spectrum (from 10 kHz. digitized data), and energy measure, a spectral change measure, and the five best results of a spectral match function (from a set of 30 to 40 phonetic targets). After application of the phonetic, phonological, and prosodic rules, overall recognition accuracy of 61.7% is achieved. In broad classes, accuracy is 88.6%. Segmentation results are also reported: 6.9% missing; 10.5% extra.

A second stage has been added to this system, consisting of a dictionary, statistical language model, and probabilistic match. Sentence level accuracy is reported as 85 and word level, 98%. 8.5 minutes of speech, consisting of 6175 segments were analyzed. [Dix75b]

4.3 Other Models and Systems

4.3.1 Dragon -- Hidden Markov Process

In discussing the Dragon system [BakJK75a], we would like to make the following observation. In a system capable of utilizing all the results of the acoustic parameter recognition, raw labeling accuracy may appear to be very poor, yet the labeling routines can support extremely accurate recognition of higher level elements provided the "correct" labels are available. The higher levels must also understand the set of labels. They must be able to use the results of acoustic recognition optimally.

The recognition statistics presented later for the BAK distance metric (which uses ZCC parameters that have been amplitude normalized) represent the primative decision rule performance in Dragon. 88 templates for 33 phone-like labels were found using the clustering algorithm described in chapter 7, and Dragon was run on a separate set of test

utterances. Very poor results were obtained. However, it was observed that a single critical error in one word could abort the entire correct utterance, because of Dragon's Markov chain representation and certain unconstrained aspects of the grammar. It was imperative, therefore, that the lexicon (in which all the phonological knowledge is encoded) include all reasonable realizations of each word so as not to be a source of a critical rejection. This was accomplished by studying the low level recognition stream of words in their occurences in the training data if the word caused a problem in the test data. Essentially, the lexicon was being trained, or rather, word-specific phonological knowledge was being acquired. Although only the training data was used to develop this knowledge, the results on separate test sentences from a fairly unconstrained grammar and moderate sized lexicon (250 words) exceeded the 95% word recognition level. One speaker was used for all these sentences.

This experiment serves as an existence proof, then, that correct machine recognition of continuous speech can be based upon what researchers have traditionally considered low performance at the acoustic-phonetic level. It also points out the need to include aspects other than first choice statistics in any analysis of labeling performance.

4.3.2 Dynamic programming

Dragon has no separate segmentation process; rather, the probability of each word in every time interval is carried through the internal representation. A similar, non-segmenting approach used by Itakura for isolated word recognition [Ita75, Ich73] is the Dynamic Programming model. In this model, stored parametric representations of an entire word or phrase are compared against the input. Time justification is accomplished by a dynamic warping of time, within certain limits, so as to optimize an overall pattern match score. A primitive parameter pattern match rule is still needed, to be applied at regular intervals (15 ms.) and Itakura introduces the minimum prediction residual -- the log likelihood ratio of one interval of the signal being predicted by the LPCs derived from the corresponding template interval.

The results of extensive experimentation with this system were 97.3% correct

recognition (1.65% rejection) over 2000 test utterances compared against 200 template utterances. The utterances were Japanese geographical names of 0.6 seconds average duration. [Ita75]

We mention this program even though it differs in many ways from the task and structure of the continuous speech understanding systems. It, like Dragon, demonstrates the power of a simple, uniform model for applying the results of parametric level recognition, and it presents an interesting use of the LPC model to directly estimate similarity of acoustic signals. It is not clear whether this approach could be extended to continuous speech, or even whether results would be good over a different set of utterances. But this clearly fills in one more point in the space of current technology available for certain speech understanding tasks.

4.3.3 Other Efforts

Two other efforts deserve at least brief mention. Hess [Hes74] reports a pitch-synchronous approach to parameter extraction. With pitch synchronous non-harmonic analysis (apparently similar to some of the LPC methods) he is able to do careful formant tracking. Segmentation produces alternating steady and transition regions with labeling accuracy reported at 85% to 90% over a set of 24 labels. (These are results of testing on the training corpus, about 1700 segments.)

Haskins Laboratories [Mer75] has been developing strategies for parametric recognition by studying human protocols on certain spectrogram reading tasks. From this they have developed Phonetic-Context Controlled methods for segmentation and labeling. The results of human performance for spectrogram reading include location of reference words in similar contexts and in different contexts and phoneme identification without reference spectrograms available. [Coo74] While these showed 70% to 80% reference word identification, and about 70% phoneme identification, correct words were found only 50% in the third experiment. In addition, the language used was English (presumably the language most familiar to the subjects) and thus use of higher level knowledge could

hardly have been avoided. Indeed, acquiring this knowledge was a large part of the intent of the experiments.

4.4 Human Performance

The implications of human performance to designers of speech understanding systems have often been misunderstood. Shockey and Reddy [Sho74a] point out two phenomena found in the discussions of human performance and computer speech recognition. The over-expectation phenomenon results in a tendency to expect too much from machine recognition at certain levels (such as the parametric analyses studied here) because humans seem to do so well. Under-expectation occurs because of the poor results in the past, and has led a very great reliance upon lexical, syntactic, and semantic constraints. The suggestion is that a balance should be struck. At the acoustic-phonetic level, one ought not expect recognition of things that are not there (acoustically). Neither should one forgive not recognizing things that are present in the acoustic parameters. The problem in performance analysis at this level is, therefore, to determine just what is present. This appears to be a strong plea for an acoustic approach to segmentation and labeling, leaving phonology and phoneme recognition to other knowledge sources. This approach is supported by the results of the experiments with connected speech transcriptions of unfamiliar languages reported by Shockey and Reddy. These results may serve as another point in the performance framework with which one may view the work at this level.

Fifty short utterances in 11 languages were recorded. Ten of the languages were unfamiliar at all levels (even the phonological level) to the subjects. Correct identification of phonetic elements was measured for transcriptions made by the subjects, who were expert at phonetic transcription techniques. The stimuli were auditory speech, spectrograms, or waveforms. Accuracy of identification into 70, 14, and 5 classes was measured. Auditory input supported results of 56%, 66%, and 78% for the three sets of classes. Spectrogram and waveform were both very similar at about 24%, 46%, and 67%.

In the light of reported vowel identification results of some speech understanding systems, it is interesting to note that the computers are doing better than phoneticians working with spectrogram or waveform data and about as well as auditory perception. When vowel performance was compressed into 6 overlapping classes: high, mid, low, front, center, and back the results were about 66%, 44%, and 46% for the three presentation media, respectively.

A different approach to studying human performance on spectral parameters was taken by Klatt and Stevens [Kla72]. Here spectrograms of unknown English sentences from an unknown but fairly simple grammar and vocabulary were used in transcription and word identification experiments. The object was to study the methods of search, particularly at the lexical lookup interface. Total segment identification performance was reported at 33% correct, 40% partially correct, 17% in error, and 10% omitted segments. Vowel and consonant identifications were each similar with the exception of a much lower omission rate for vowels. An interesting result of the study of lexical interactions was that most of the searches initiated did not yield the correct word. However, after extended interaction, (and obviously application of some higher level knowledge) word recognition was improved to 96%.

There are a large number of other experiments which deal with issues of human perception of speech, although they are often intended to reveal aspects somewhat irrelevant to this dissertation, such as the perception of altered speech, superimposed sounds, context, dialect, speed, or stress. We am concerned here with the much more basic problem of robust acoustic identification and segmentation. However, human perception performance does serve as another point in the space of speech understanding systems, as well as an existence (or, sometimes non-existence) argument for certain acoustic-phonetic correlations. Ladefoged [Lad69], in discussing perception of vowel quality, points out a great deal of ambiguity in the way vowels are perceived and described by phoneticians and linguists that seems to invalidate much of the detailed analysis of vowel quality as being descriptive of real physical phenomena. Dealing with

specific acoustic gestures (with whatever phonetic correlates they may possess) is more reasonable than attempting to do phonetic or phonemic recognition. Structures and rules such as Markov chains, Sniffers (SDC), dynamic programs, etc., may have little to do with the kinds of things traditional linguists have studied, yet may be better oriented to computer understanding of speech. It is not at all unreasonable that in the context of different systems (from humans) and different acoustic representations, the optimal phonetic classes and types of segments may not be the same as those defined by traditional phonetics.

4.5 Summary

In this chapter, we have presented some of the results and relevent aspects of a variety of phone recognition components from a number of current speech systems. Direct comparisons are very difficult, although some have been made. [Hie75] There is wide variety in the types of knowledge used, the representational level for output, and the expectations and system demands which characterize these recognizers.

From these descriptions it may be possible to form a picture of the uses to which parametric level analyses will be put. It is impossible, however, to form an accurate estimate of the performance expected at the state of the art today. Depending upon conditions of testing, knowledge sources used, level of represention, etc., the reported accuracies may vary a great deal. There are a few existence proofs, even so, that may indicate that the state of the art is approaching a point where general connected speech understanding will be possible. In a number of limited domains, very high accuracy and even real-time responsmay be achieved.

Chapter 5

Segmentation Procedures

This chapter presents a method for segmenting speech into acoustically uniform segments. The algorithm is relatively independent of the choice of parametric representation (except for the use of one amplitude parameter specifically as such). The basic operation is the application of a distance metric to patterns adjacent in time. A special metric has been devised for this purpose. A second aspect of the method is that it employs a number of different decision functions.

5.1 Role of Segmentation

There is a very strong interdependence between the performance of segmentation and of labeling at the acoustic level and, indeed, at higher levels. The dependence operates (in one sense) because of the need for labeling to be performed on the "target" areas of the signal -- those portions of the phonetic gestures' duration in which the articulators are closest to their intended positions, during which the excitation source is most stressed and steadiest, and during which coarticulation effect may be minimized. In the opposite sense, interaction occurs because segmentation is highly dependent upon context. As simple a cue to segmenting as amplitude or energy in the signal is much less meaningful during strong fricatives, such as /s/ and /z/, than during most voiced segments. In the former cases, amplitude may carry no phonetic information at all. In the latter, it often signals some important change, such as a vowel/glide juncture, and may be the only robust signal of some pathological cases. Thus the information gained in classifying the phonetic context of an acoustic change in the signal can be very important in determining its relevance to phonetic changes.

Different approaches have been taken in a number of speech recognition programs to deal with this interaction. The Hearsay I system [Erm74b] attempted to do labeling first, by placing acoustic labels on the signal at regular, short intervals. The string of

acoustic labels was then smoothed, and points of change in the nature of the labels were located. Additional information, such as amplitude and the location of amplitude peaks and dips was also used to improve this label-driven segmentation. After changes in the gross signal type (voiced, fricated, silence) were located, voiced segments were further subdivided using the amplitude clues.

Another, very different, approach [Ita75] is to ignore segmentation entirely as a separable process. Rather, a dynamic programming problem is postulated for solution. In this problem, short time windows at regular intervals are matched by means of some general matching function (labeling), the two intervals for each match being taken from the input signal and from a stored template. Within certain constraints, time may be warped so that different regions of the unknown are matched with different regions of the template (segmentation). The solution to the dynamic program is the time warping which "best" aligns the two signals; the solution also provides a rating of their match to be compared with those of other templates.

Both of these methods, as well as other, similar approaches have been successful in limited speech understanding systems. However, acoustic level input to a speech understanding system of more general scope must take another point of view. Crude segmentation may be sufficient for tasks with small vocabularies and high degrees of syntactic and semantic support. No segmentation at all is a possible approach in recognition of short utterances (single words) where the advantages of a uniform algorithm are not overweighed by an excess of computational and storage requirements nor by the chance of significantly wrong paths being taken in the "search" for an appropriate time warping. However, in dealing with continuous speech over weakly constrained subsets of natural language, the difference between two semantic outputs may rest upon locating a few glides and nasals as separate entities from the surrounding voiced context, or may rest in ignoring large intervals of irrelevant signal (silences, offglides, etc.). In fact, an accurate segmentation is more than half the battle in many cases, since an accurate account of the number and general nature of the phonetic elements can greatly reduce higher level searches.

5.2 Present Segmentation Method

For such reasons, we have been investigating the feasibility of segmentation-first machine transcription. In addition, such machine-produced segments provide a realistic test of labeling performance for the different parametric representations. In actual practice, a parametrization of the input signal must support both processes of segmentation and labeling. If critical errors are made in either, the total performance will suffer. Because the choice of representation is a very open question, we have attempted to develop algorithms which are, to a large degree, parametrization-independent. Although certain prosodic parameters are necessarily assumed and relied upon within the algorithms, the general approach has been to develop segmentation and labelling at a level where various representations are treated in a uniform manner. Thus, the programs that result are also useful research tools for this comparative study.

5.2.1 Detecting Change

The basic concept of segmentation is very similar to the well known signal detection problem. (A more detailed discussion of this model is presented in the next chapter.) In this case however, the signal to be detected is "significant" change. Given some function of time which measures change and which can operate within a small time window on the input signal, we can postulate two distributions of values for that function: those correlated with times of acoustic or phonetic change, and those correlated with times of no relevant change. Unfortunately, the form of these distributions is not available a priori, although we may not care about their form. So long as a change measuring function is available which produces significantly different (higher) values at just those times when changes are occurring in the phonetic state of the signal, the segmentation problem may be solved. The actual distributions are useful, in signal detection, for choosing optimal thresholds for the detector. Given the costs associated with type I and type II errors, one can balance the decision. However, the costs are not really known here, since they may

rest upon much higher level considerations (the phonetic similarity of two semantically different words, for example). We would hope to learn enough about the change function from empirical analysis for our purposes.

From the previous discussion, it appears unlikely that any one measure of change will suffice for the segmentation of continuous speech since the changes are context dependent. To look at the problem from a pattern space point of view, different regions of the pattern space will change in different ways. The intra-class distances of samples of /s/, for example, have been found in some parametric representations to be much greater than those of /n/, and they vary in different elements of the patterns. We can take this view further and rely upon a distance measure in the pattern space to compare neighboring (in time) patterns of the input signal and to rate the likelihood of a phonetically significant change. In addition to merely rating the likelihood of change, we wish to locate it in time as accurately as possible. If the resolution with which we look at the signal is fine enough, we may assume that neighboring high values in the distance-of-neighbors function relate to the same segment boundary, and we ought, thus, to choose the time of the highest value, the local peak.

5.2.2 Multiple Decision Algorithm

Since the following discussion goes into the details of the segmentation program, the reader may wish to skip the rest of this section after looking at figure 5.1 for a general idea of the method.

The first approach [GolH74] was to employ a single decision function which combined measures of both short and long duration changes. This proved to be too inflexible. No simple threshold could be found on such a composite function to separate change-related from non-change-related peaks. If this were because the distributions of peaks both at segment boundaries and not at boundaries were identical, then the function would be useless. However, we found some value in the function if it was employed with different thresholds, and with varying resolutions in time, during different portions of the

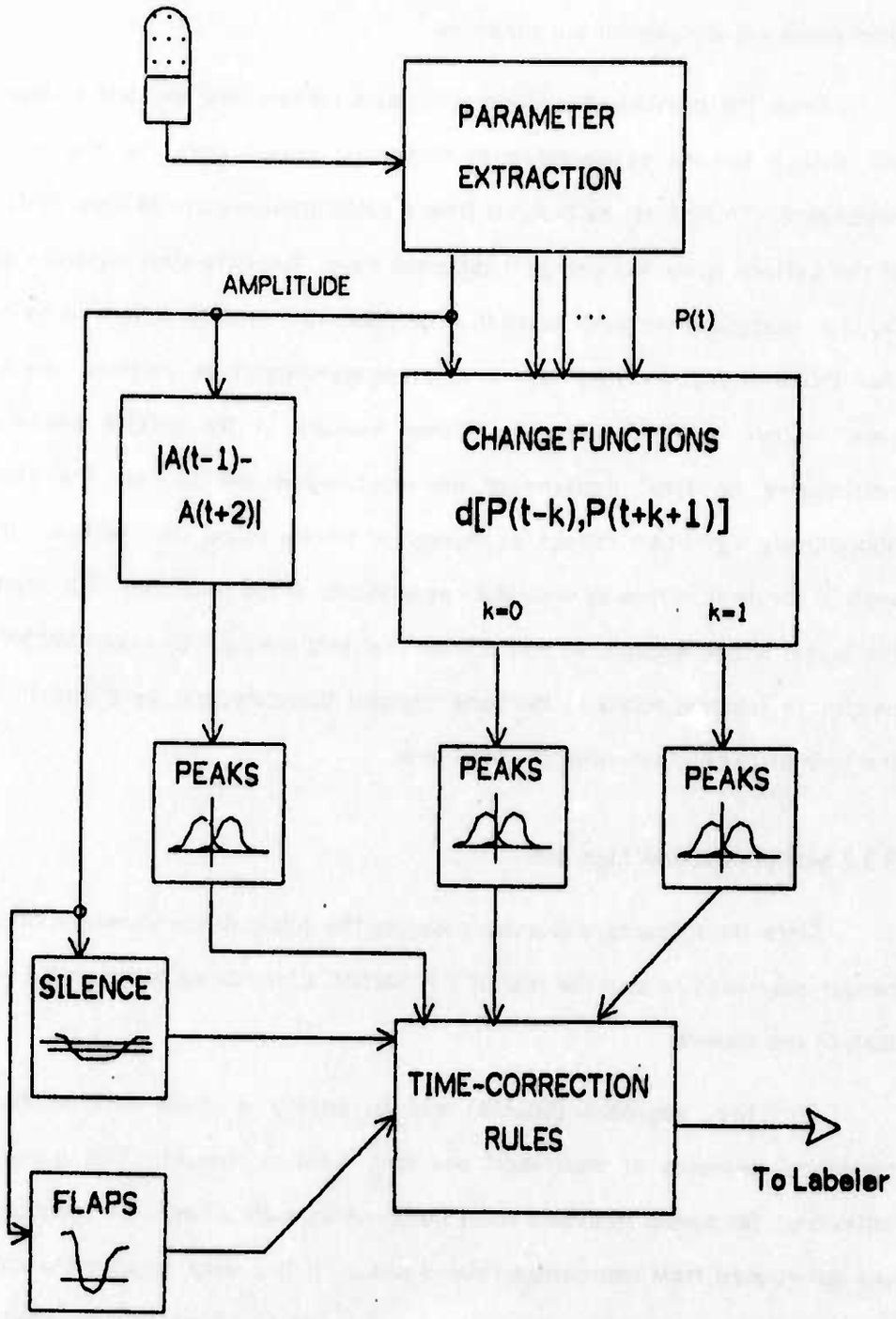


Figure 5.1: Segmentation Program

signal. It appeared that the distributions were overlapping significantly because they were composites of different kinds of phenomena.

The current segmenter, therefore, first attempts to locate points of major change in the signal source and in amplitude. Then less robust functions are used to segment within these broadly separated regions. Additionally, some correction rules are applied to adjust certain cases where two functions do not quite coincide in the time-placement of segment boundaries, but where they are both clearly responding to the same signal change. The following discussion describes the philosophy of combining evidence from a number of segmentation functions, the decision functions themselves, and the correction rules. Then a discussion is presented of the training of thresholds and the rating of segment boundaries.

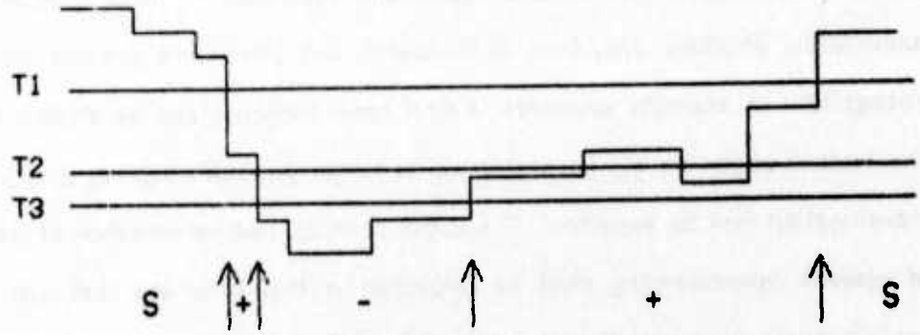


Figure 5.2: Speech/Silence Detection

As the first stage (see figure 5.2), speech is separated from silence segments by locating those times when the amplitude parameter has dropped below a threshold, T1. If the amplitude also falls below a second threshold, T2, at some point in the segment, it is accepted as a silence region. (The acquisition of values for T1 and T2, as well as thresholds used by other levels, will be discussed later.) The amplitude parameter plays a special role in the segmenter. Because of limitations in the accuracy of digitization, as well as inherent shortcomings of the methods, many parametric representations are unreliable at low amplitudes. It is important that the regions of the signal be located where analysis can be done with greatest confidence. Moreover, amplitude carries important segmentation information which ought not to be overlooked when, for reasons of normalization, it is often removed from the parameters.

Within the regions identified as silences, a third low threshold T3 is applied from either end of the segment to separate onset and offglide regions, or very low amplitude nasals or fricatives, from true silence. These regions may belong "phonetically" with neighboring, higher amplitude segments, but have been divided by the application of threshold T1 from the main part of the speech region. It is a serious question whether this sort of segmentation is an error for a mechanical segmenter. Our goal in machine segmentation ought not to be to duplicate hand segmentation of phonemes, but rather to isolate those locations in the signal which will best support labeling and will provide as accurate and reliable a map of the acoustic reality of the signal as possible. Higher level rules in the speech understanding system can then deal much better with the problem of fitting phonemes to the acoustic labels and segments. Therefore, in isolating the low amplitudes, offglides, *etc.*, from both speech and silence, we provide for their possible recognition as separate segments (a final nasal perhaps), and we ensure that labeling of the speech segments will be performed as higher amplitude signals, where more accurate classification may be expected. The above premise, that performance of the acoustic level of speech understanding must be measured at that level and that one cannot expect certain kinds of recognition behavior from the simple, local algorithms one encounters there, is very central to our approach in this research. We will meet it again in other contexts.

There is one other speech/silence boundary phenomenon which must be dealt with at an intermediate level. Flaps will not, in general, drop in amplitude to a level where the silence detection described above can detect them. However, the flap does have a very particular kind of amplitude contour. (Figure 5.3) Thus, in the speech portions, short periods of amplitude below a threshold T_l together with preceding drop T_p and succeeding rise T_s are isolated as flaps. Only durations of 10 or 20 ms. are considered. We have observed that flaps of longer duration are adequately detected by the other functions of the segmenter. Moreover, it is a point of phonological debate when an intravocalic stop consonant is really a flap. Thus the program only labels as flaps such stops of 20 ms. duration or less. (The coarseness of the resolution, 10 ms., may allow stops of 30 ms. to be detected.)

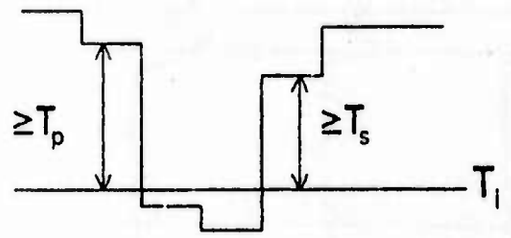


Figure 5.3: Flap Detection by Amplitude Contour

From this point on, a number of detection functions are used. Figure 5.4 displays some typical ones, as well as the hand and machine detected boundaries.

At the second stage, the speech regions (between silences and/or flaps) are subject to the first decision function, V_f. This may be any acoustic distance function. In this case it is a vote-for-change function computing in the following manner: the difference of successive parameter values, $|P_i(t)-P_i(t-1)|$, is composed with a threshold R_i, where i=1...(number of parameters). If the threshold is exceeded, a vote of 1 is accumulated. The acquisition of R_i will be discussed later. This function will peak at a time when the parametric representation is changing in a number of its elements. The local peaks above threshold T_f are considered strong candidates for boundaries of significant change in signal type.

At this point a third decision function, V_a, is applied. This is the magnitude of the change in the amplitude parameter between t+1 and t-2. This larger duration measure was adopted because the object of this level is to find fairly major boundaries between voiced segments, such as vowel/nasal junctions and even certain vowel/stop boundaries which escape the previous function because the patterns are similar in overall type. It was found that using shorter duration amplitude changes introduced too many spurious decisions while a longer span would confuse amplitude changes of a gradual nature with shifts that signal phonetic change.

By now, boundaries of three different types have been found, major silence/speech

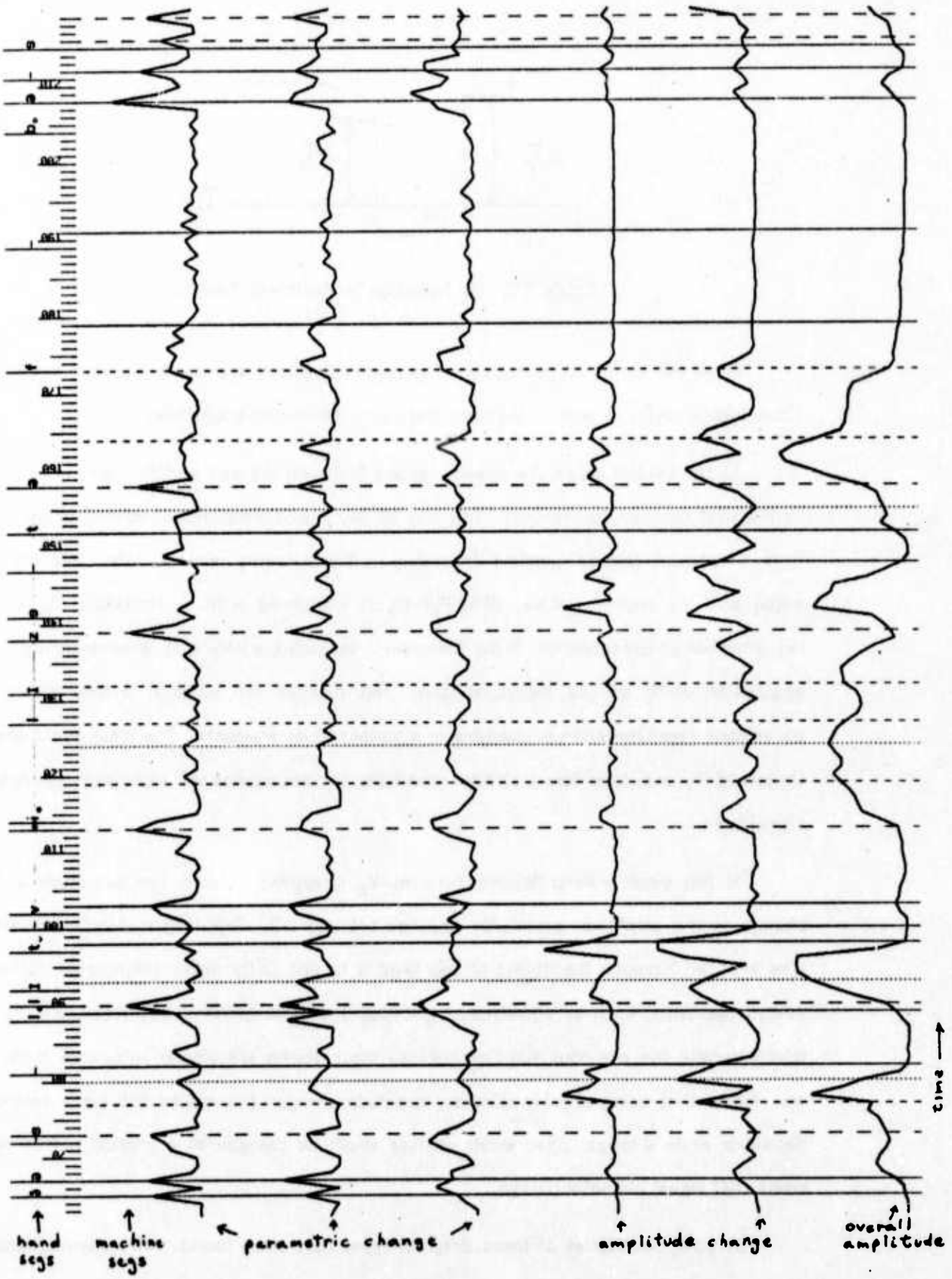


Figure 5.4: Some Detection Functions (and Amplitude)

junctures, points of major change in signal types, and significant changes of amplitude within speech segments. The final decision level deals with slow changing boundaries of the kind encountered in vowel/glide junctures or in diphthongs.

By the time segmentation has proceeded to this fourth function, we may assume that all the obvious boundaries have been detected. Therefore one ought to take a general look at the input patterns in order to detect changes which persist for a length of time. In addition, some errors of omission from higher levels may be corrected here. Such a situation as an /l//z/ boundary, where, because of the high third formant of the /l/ and the voicing of the /z/, as well as similar amplitude levels, an obviously important boundary is missed, will be corrected by a function that is more sensitive to the acoustic change in question. Consequently a difference threshold, S_i, is applied to each of the pattern channels and a vote sum, V_s, is taken similarly to that of level two.

The differences above are taken between the input patterns at times t+1 and t-2 for reasons similar to those concerning the amplitude difference function V_a. This function, V_s, is intended to detect slow changes in the signal, the windowing gives a 30 ms. span to the observation of change.

There are a number of situations where two or more of these functions respond to the same change phenomenon yet locate it at slightly different times. This may occur because the different functions are sensitive to different portions of the pattern, or because of scaling or windowing considerations of the particular parametric representation. Therefore, after the various levels mark the boundaries, correction rules are applied. For example, speech silence (level one) boundaries may be corrected 10 ms. to the location of level two boundaries in situations where speech is going to silence. (In speech onset cases, one cannot afford to miss short burst segments by correcting this way, but in off-glide cases, there are no such short segments.) V_a and V_s boundaries, being responses to phenomena spread over 40 ms., may differ by 20 ms. and be corrected to the point in this region where a general difference vote function is highest. These correction rules have been developed in an empirical manner. However, we have

attempted to maintain a sense of justification for any such rule. The limitations of a 10 ms. time-resolution also justify the assumption that errors will consistently be made in the time location of more gradually changing boundaries.

Issue might be taken with the use of thresholds in so much of this process. One can never rely completely upon thresholds to perform under varying conditions. However the data to be used in this study has been accumulated and analyzed under fairly consistent conditions, with regard to noise, gain, and microphone characteristics. This assumption of consistent data should not affect the more general applicability of these methods, since amplitude and spectral normalization techniques are fairly commonly available. For example, Itakura gives a method of normalizing the long-term spectra which essentially models the transducer characteristics with a two variable linear equation. [Ita75]

5.2.3 Training

At this point we should discuss the problem of training-- that is acquiring values for the various thresholds and weights mentioned above. It is important to the goal of this comparison that these values be acquired in a uniform fashion for all parametric representations. Uniformity was equally important in order to maintain some detachment from the initial test data upon which the segmentation program was developed.

The functions V_f and V_s depend upon vectors of thresholds. These have been derived empirically from a corpus of training data which is segmented and labeled by hand. It consists of 27 separate utterances of continuous speech of about 3 or 4 seconds each. Given a parametric representation, $P_i(t)$, where time, t, advances one unit each 10 ms., all the times at which hand segmented boundaries occur (within the 10 ms. resolution) are considered. At each such time, t, the differences, $d_i^1=|P_i(t)-P_i(t-1)|$ and $d_i^2=|P_i(t+1)-P_i(t-2)|$ are calculated. The greatest d_i^1 and d_i^2 are then collected by assigning the threshold R_i to be the least such d_i^1 and S_i the least d_i^2. Although a large amount of data may be thrown away, this selection ensures that only the most significant parameters in the representation are given low vote thresholds. Clearly, however, this method is sensitive

to any errors in the hand segmentation, and great care has been taken with the training corpus especially.

The calculation of V_f, V_a, and V_s is thus dependent upon the training corpus only. The next step is to describe a method for finding the detection thresholds, T_f, T_a, and T_s used to accept or reject peaks in these functions as signifying segment boundaries. Signal detection theory can help us derive optimal detection thresholds from statistical analysis of the stimuli. A typical detection model postulates a univariate measure, L, of the stimulus that relates to the two events, signal-plus-noise, S, and noise-only, N. Very often, a convenient measure is the ratio Pr{S|input}/Pr{N|input}. It is further assumed that the distributions of L|s and L|N are distinct, very often identical except for different means. In standard studies of signal detection, L cannot be accessed. Rather the forms of the distribution are assumed, and the actual performance of the detector is used to measure the difference in distribution means. The basic model shows that performance, as measured by both Pr {detection |S} and Pr {detection|N}, can be optimized for any set of costs. As we shall see in the following chapter, when we discuss the signal detection model further, the choice of threshold is really an arbitrary tuning device. The choice of decision function has the primary effect on performance.

In the segmentation program described above, the signal (boundary) measure is known, and we may collect distributions empirically. Therefore histograms were collected over a corpus of training utterances to estimate the distributions Pr {L|S} and Pr {L|N}. Since the measures V_f, V_a, and V_s are coarse in time, a peak was considered to correlate with a boundary (S) if it was within 10 ms. of the hand marked boundary. The histograms of local peak values within 10 ms. of segmental boundaries estimate Pr {L|S} for L=V_f, V_a, or V_s.

At this point, we had to decide what costs to assign to errors and what confidence to place in the histograms. The distributions all had the same approximate shape, although somewhat different apparent variances and different means. Since the means of the populations could be estimated with the most confidence, the thresholds were chosen to

be halfway between the two distribution means. This corresponds to the model with equal variances, same distribution, and equal costs-- a simplification of the observed situation. The figure (5.5) shows some empirical distributions for the SPG representation and a possible threshold value for one of the three decision functions.

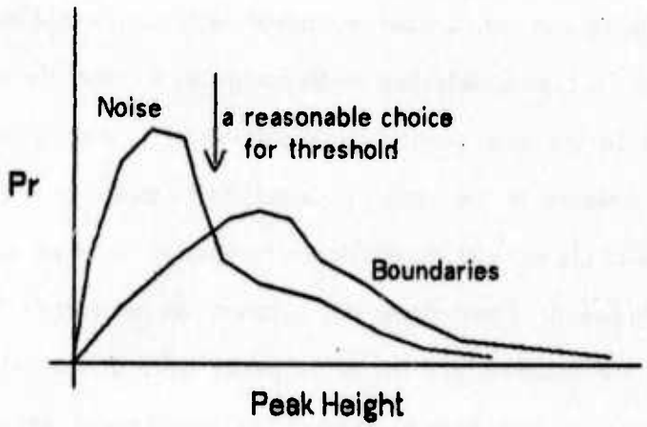


Figure 5.5: Choosing Thresholds

Finally, the acquisition of the thresholds for silence and flap detection has been more ad hoc in nature. Amplitude values were collected over the training corpus for silence segments within utterances (which might be more noisy than inter-utterance silences). The mean and standard deviation were computed and T1 (the boundary locating threshold) is assigned mean and standard deviation. T2 (silence verification) is mean, and T3 (low amplitude segment location) is the mean of amplitude readings over silence, /b/, /d/, and /g/ segments, tending to be slightly higher.

The flap thresholds were chosen by observing the behavior of the amplitude function in a small number of cases. The flap is a rare enough phenomenon so that it is difficult to collect adequate statistical information about it. Moreover, since flaps are one specific performance and occur in limited contexts, it is unlikely that the observed patterns were not representative.

5.3 Summary

We presented the outline of a segmentation procedure. The details provided are of interest only to those who may be involved with speech segmentation. However, the general scheme should be of interest to others. Speech may be separated into broad class regions first -- silence and speech, or if voicing can be detected, silence, sonorant, and fricated speech. These regions require different types of segentation activity. However, most of that activity can be based upon a simple detection process with very good results. We have introduced some rather *ad hoc* methods of training, which still work well. One reason for this, is that the choice of detection thresholds is guided as much by the requirements of the speech systemas regards the missing-extra trade-off. Thus, the tuning of thresholds is a job for the system designer.

Chapter 6

Segmentation Performance

In this chapter, we will deal with some of the problems involved in evaluating segmentation accuracy. The difficulty of obtaining a correct standard for comparison with the machine determined boundaries, the nature of various types of discrepancies between the machine boundaries and the standard, and a normalized measure of accuracy of segment detection which is derived from signal detection theory will all be discussed. We will also present the results of segmenting continuous speech employing the four parametric representations chosen and described earlier, ZCC, SPG, ASA, and ACS†. Finally, some experiences with the use of the segmentation algorithms, as well as case analyses of particular interesting "errors" are included.

6.1 Evaluating Segmentation Errors

Errors in segmentation of continuous speech must be considered in the light of reasonable expectations. If a segment boundary is actually indicated in the acoustic input, then it ought to be detected. In a like manner, indications of boundary-like change that do not correspond to phonetic boundaries should be accepted as legitimate results of segmenting without higher level knowledge sources. Alternatively, boundaries that are not indicated by some change in the acoustic signal (such as the nasal/nasal juncture in /some milk/) should obviously not be expected from a segmentation procedure which only examines the acoustic parameters. Another, equally important, consideration in evaluating a segmenter is the effect of its errors on the overall speech understanding system. Some speech understanders are more sensitive to missing segments (boundaries) while others will handle these well, but become overloaded if too many extra segments are "detected".

If the above statements are taken as a definition (or description) of the kind of discrepancy between standard and test transcriptions which we wish to consider an error,

† using Itakura's distance measure

then a problem arises with respect to the usual structure of hand transcriptions of continuous speech. The transcription process is very often one of successive refinement and, consequently, descent through the "levels" of linguistic elements. First the words are considered, and approximately fit to strong features in the signal. Then an attempt is made to fit the standard phonemic spellings of each word, with corrections being made whenever phonological or phonetic interactions are detected. Finally, if the transcription is at a sufficiently fine level, the short-term acoustic nature of the signal is used to infer sub-phonemic segments (sometimes due to co-articulation effects) as well as phenomena which might not be explained by accepted models of speaker or language, but which can be justified by the acoustic data. The transcription may, therefore, have segment boundaries which are justified by strong or weak acoustic cues, by phonemic expectation, by phonological or phonetic rules, or by any combination of these factors. Thus, when the machine segmentation misses such a boundary, we must determine what kind of a boundary it is to determine whether an error has been made. Likewise, when we have marked an extra segment error, we must be sure that there is really no justification for a boundary at that point in the utterance. We are often not in control of the source of the hand transcriptions used for evaluation standards. These may also be used for other purposes, and, since they are costly to acquire, may have to include the kinds of information just mentioned.

In a paper describing their segmentation and classification evaluation system [Sil75], Silverman and Dixon discuss the difficulty of acquiring standard (referent) transcriptions of continuous speech. This difficulty is compounded by having different sets of phonemic/phonetic elements. For example: /ch/ versus /t//sh/, /t/ versus /-//t<burst>/, or /el/ versus /e//l/. Their philosophy appears to be similar to ours in that they consider sub-phonemic segments when evaluating, but only insist that phonemic segment not be missed. They also collect statistics on missed and extra segments, with the addition of separate statistics on misplaced boundaries. These are defined by specifying alternate transition and steady state regions and declaring segments to be properly placed if they fall within the appropriate region. Segmentation errors can be reported individually for each phonetic label.

In her dissertation involving a new acoustic analysis method [BakJM75], Baker evaluates the performance of five segmentation programs on a small set of sentences (5 sentences -- about 200 segments). Her evaluation was performed by hand and extremely careful measurements were made of the mislocations in time for various classes of speech sounds, as well as the usual missing and extra counts. Her philosophy concerning what boundaries are important agrees very well with ours, as well as Dixon and Silverman's. Thus we may compare all the segmentation results from these two sources with our own and each other. The following caveat must be offered, however. A number of the routines tested by Baker were in the early stages of development. In addition, the quantities of data are very small to draw any far reaching conclusions. Finally, some way of normalizing for different decision criteria, which lead to large amounts of trade-off between extra and missing segments, must be applied. The following section on a signal detection model will provide such a normalization. The Results section of this chapter presents both our own segmentation performance and what we feel is an accurate interpretation of the other reported performances.

One way to rectify the problem of acquiring good standard segments is to use two transcriptions (possibly derived from one another). One should contain all the segments that might ever be justified (the union of the various descriptive levels). The other should contain the segments that are both acoustically and phonemically justified (the intersection). A set of segmentation boundaries, M, are evaluated by comparison with these two standard transcriptions, H1 and H2, in the following way. (See figure 6.1)

```
Hand 1    ..*..*.........*.............*.....*....*..
Hand 2    ..*..*.*.......*.............*....*.*....*..
Machine   ......*.....*.....*....*.............*.*....*..
Error     ..M.........X.........X.......X..............
```

Figure 6.1: Evaluating Segmentation

Given a margin of admissible error†, an EXTRA segment boundary is recorded if there exists a boundary in M at time t but none in either H1 or H2 in the interval [t-margin,t+margin]. Similarly, a MISSING segment is recorded if there are boundaries in H1 at t1 and H2 at t2, $|t1-t2| \leq$ margin, and there is none in M in the interval [min(t1,t2)-margin,max(t1,t2)+margin]. Obviously, when H1=H2 this becomes an obvious comparison of two sets of segment boundaries.

This technique has allowed us to use hand transcriptions which are actually rather inappropriate for segmentation performance evaluation. However, there are, in any corpus of continuous speech, a number of segments which may be best described as transients. Unless the hand transcription has some indication of these, their detection by a machine segmenter will show up as extra segment boundary errors. While many such transient segments are indicated, a careful inspection of the extra segment errors discovered about 300 additional segments in a corpus of about 1000 primary segments which were not originally marked in even the careful hand segmentation we have available for this corpus. The number of EXTRA segments is usually not as critical to system performance as the number of MISSING segments. However, each serves at least as a counter-measure to the other. We will observe, in the following section on signal detection that the response characteristics of a subject in a series of detection trials will trade off correct positive with correct negative responses. Similarly, one can tune the segmentation algorithms to deliver many more segments, thereby increasing the number of EXTRA and decreasing the number of MISSING segment boundaries. So the two statistics must be considered in conjunction to determine the detectability of segment boundaries.

An additional test was added to the evaluation routine described above. This test checks for pairs of MISSING and EXTRA errors which 1) are close together (e.g. ≤30 ms.) and 2) have no intervening hand segment boundaries indicated in H1 or H2. These pairs may be taken together as cases of misplaced boundaries since, if there were any other significant phonetic change in that region, it would be indicated in the hand segmentations.

---

† We are using a margin of 10ms.

This correction to the statistics is made after the entire utterance is examined for missing and extra boundaries.

A final point concerning evaluation of segmentation is the place occupied by hand evaluation. We have had to rely, for some of the evaluation fidelity, upon hand inspection of the waveforms. This has been necessary to determine how critical the errors may be and whether, in fact, they are errors at all, rather than mistakes in the standard transcription or cases of non-acoustic boundaries with no acoustic correlates for missing segment errors, or the reverse for extra segment errors. This is especially important because of the low percentage of boundaries which are missed, and the consequently large effect of each incorrect evaluation decision. The MISSING segment errors are typed as Type 2 if they are critical to recognition or if the acoustic cues are clearly present and there is a significant phonetic juncture in the area. Type 1 represents less critical boundaries. Often the lack of a segment may be explained by some reasonable phonetic theory of the speaker's performance. In other cases the boundary has been detected, but at a different point in time. Often, with slow transitions, the exact time location of boundaries is impossible. Type 0 errors are usually cases where the standard transcription is in error or is not acoustic in nature. These are boundaries which we cannot expect the segmenter to detect. In the case of extra segment boundaries, we have marked as type 1 those boundaries which appear to have no acoustic validity upon inspecting the waveform. Type 0 EXTRAs are, again, places where the standard transcription is a non-acoustic description, at best. The cases presented in Appendix S1 may serve to indicate the kind of problems one will encounter in dealing with continuous speech - no matter how carefully one thinks the standard transcription has been prepared. They will also give an idea of the performance of the segmentation algorithms which we are using. A subset of the speech corpus, presented as oscillograms, with the standard transcription, is included as Appendix S2. Marked on this display are the segmentation errors for one run (SPG parameters, speaker CC) and their type *vis.* the previous discussion.

6.2 A Signal Detection Measure

In this section, we will present the basic mathematical model of Signal Detection Theory and discuss its application to the evaluation of machine segmentation of speech. The parameter d' is a measure of the inherent detectability of the signal versus noise which is independent of any decision process application of costs or a priori likelihoods. We have empirical evidence that the model fits the actual performance of our segmentation algorithms for at least one parametric representation over a wide range of detection thresholds.

The theory of Signal Detection as formulated by Tanner et. al. [Tan64, Lic64] is primarily applied to detection trials which may be considered similar to the segmentation process. Detection trials consist of a series of responses to stimuli which may be composed of noise or of noise and some known signal -- not unlike the decision process resulting in the placement of a segment boundary based upon local information only. It is assumed that the a priori likelihoods and costs of various errors are known to a decision process which senses and possibly transforms the stimulus into some internal signal space before it yields an optimal† decision on the presence of the signal. The detector's sensory data is considered, in this model, to be reduced to a single decision parameter. A reasonable one might be the ratio of the probabilities that the input stimulus was signal+noise versus noise alone. A simple threshold on this single parameter may be placed to optimize the expected costs given a priori likelihoods, costs of misses, false alarms, etc. Figure 6.2 represents such a hypothetical internal decision parameter.

Very simply stated, the model assumes a single decision parameter, x, which may be any sensory measurement one wishes. The distribution of x values for the two types of stimuli, signal+noise and noise alone, are assumed to be normal (with equal variance in the simplest version of the model). Their means differ by d' times the standard deviation.

---

† in a decision theoretic sense, given the a priori knowledge about the test

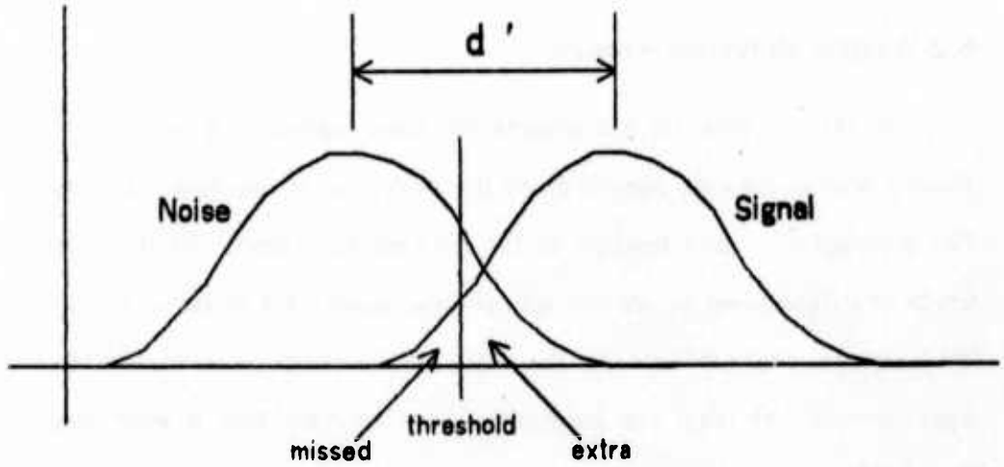


Figure 6.2: Signal Detection Model

Rates of "hit" and "false alarm" -- Pr{accept|signal} and Pr{accept|noise} respectively -- are sufficient to determine the least d' for which an optimal decision process can display the observed rates. When the hit and false alarm rates are plotted against one another for a number of sets of trials where the detector's acceptance threshold has been altered, a response operator characteristic curve is obtained (see figure 6.3).

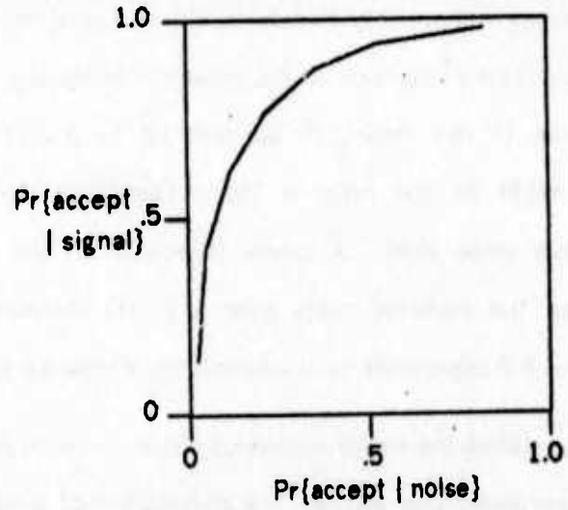


Figure 6.3: Typical ROC Plot

The theory states that the curve is totally determined by d'. When the axes of the ROC curve are transformed by the inverse function of the Normal distribution function, the

curve is approximately a straight line with slope=sigma(noise)/sigma(signal) and x-intercept=d'.

This theory has been most often applied to detection trials to provide estimates of the detectability of the signal as it appears in a human perceiver's internal sensory signal space. The estimate provided by the signal detection model may then be compared with well known properties of visual or auditory signals to provide a bound on the efficacy of the perceiver's transduction process -- the sensory channel. While the main thrust of its application is not relevant here, the signal detection model and the dimensionless measure d' can be used as a normalized measure of segment boundary detection that is relatively unaffected by adjustments in the proportion of missing versus extra segment errors. Furthermore, the d' value once estimated may be used to predict the entire response-operator characteristic.

Following the procedures shown in Egan et.al. [Ega64], a series of segmentation runs were made with the ZCC parametric representation for a set of 40 utterances (TAP) in the AP news retrieval task domain with one speaker. These runs were to investigate a range of detection responses by varying the thresholds used internally in the segmentation algorithm (see Chapter 5). The resultant error rates may be seen in figure 6.4 below plotted on inverse Normal axes. A linear least-squares fit to the points yielded a slope of 1.000 and a d' value of 2.250. We will, henceforth, assume that the simple model with equal variances gives a good estimate of the performance of the segmentation algorithm we are testing. The d' values reported below will be derived from that model.

Finally, a confidence interval was calculated for the d' statistic, under the assumption that it is approximately normally distributed for any particular experiment. The 95% interval for the SPG experiment was $\pm$.14 in d'. We shall see that this is considerably less than the differences in d' observed among parametric representations.

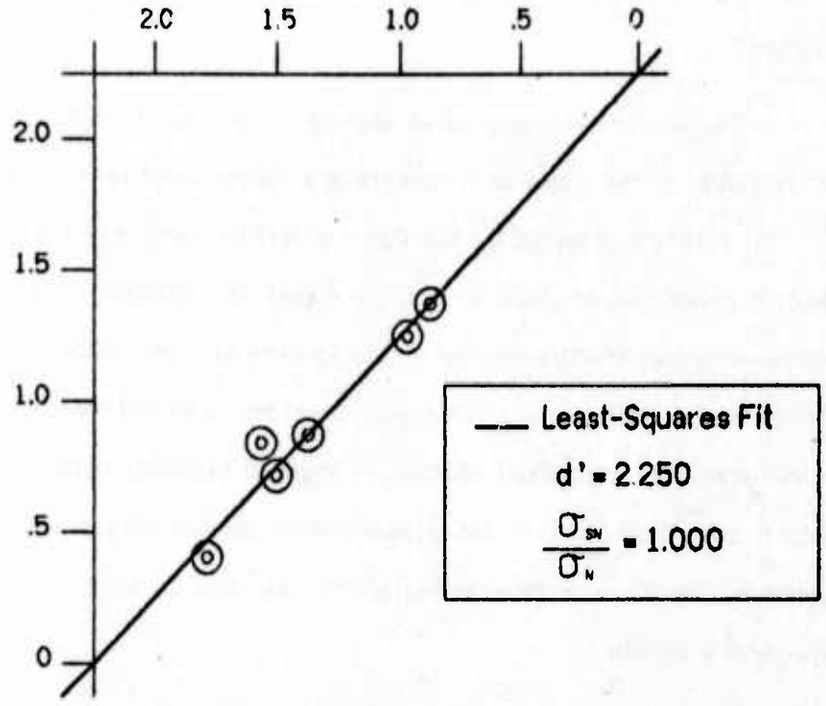


Figure 6.4: A Series of Segmentations with Different Thresholds

6.3 Results

As we mentioned in Chapter 5, the segmentation procedures used here were developed for use with the Hearsay II speech understanding system. In keeping with the philosophy of that system -- the separation of knowledge about speech into individual modules or Knowledge sources -- we employ no phonetic or phonological knowledge to correct segmentation decisions. Indeed, we do not even use the labeling information to join similarly labeled adjoining segments at this stage. It is difficult, therefore, to compare our results directly with other segmentation and classification schemes which interact closely to produce a transcription at the phonetic or phonemic level. What we have done, however, is to carefully inspect the errors made by our segmenter with the SPG

parameters as the input representation.† Then a new reference segmentation was produced by adjusting the old one to all of the type 0 errors (those cases where the hand segmentation was in error, or the machine segmentation agreed with the acoustic signal).

The following table (Figure 6.5) shows the results of segmentation for the 40 sentences from the News Retrieval task (TAP), spoken by CC (male, American).

	SPG	ACS	ASA	ZCC
Segments #				
H1	1082	1082	1082	1082
H2	1541	1541	1541	1541
M	2026	2082	2010	2290
Missing				
#	37	57	91	52
%	3.7	5.8	9.2	5.3
Extra				
#	299	391	434	681
%	27.6	36.1	40.0	63.0
	(19.4)	(25.3)	(28.1)	(44.2)
Shifted				
#	28	34	45	41
%	2.8	3.4	4.5	4.1
d'	2.38	1.93	1.58	1.29
	(2.65)	(2.24)	(1.91)	(1.77)

Figure 6.5:Segmentation Performance -- Different Parametric Representations

The first reference segmentation contains 1082 segments primarily at the phonemic level of description. The second reference contains corrections to this file, as described above, to make it more an acoustic description of the corpus. It has 1541 segments. The number of machine segments reported may be greater than the sum of this number (hand reported acoustic segments) and the number of extra boundaries. The discrepancy is merely the result of the way we evaluate segmentation by boundaries. Occasionally, two machine boundaries will fall close enough to a hand boundary to both be accepted. Such segments must, therefore, be very short, and are usually transition segments which may

† Decisions were made from inspection of the waveforms only. The results of this hand evaluation are included as Appendix S2.

be easily detected at higher levels. The number of missing boundaries (segments), divided by the number of boundaries which are included in both reference segmentations, is the missing segment error rate. The number of shifted boundaries is also divided by this number. The number of extra boundaries is divided by the number of primary segments (the size of H1 in this case). The extra segment rates in parentheses are those where division is by the number of acoustic segments (size of H2). Finally, d', as determined by the missing and extra segment error rates is calculated from the equal variance model discussed in the last section of this chapter.

In d', we can see a clear decrease in "detectability" of segment boundaries which agrees well with what we would suspect about the information content of the four representations.

A second set of results were obtained over two data sets for which no such carefully compiled hand segmentations are available. (See Figure 6.6) They can, however, be compared with the machine segmentation results just presented, as a demonstration of the robustness of the algorithms used, and thus the validity of evaluation made with them. We employed the ZCC representation for this experiment because of its availability, although any of the parametric representations would have provided just as valid results. In this evaluation, we used only the primary phonetic level hand segmentation. The results of comparison with this not necessarily acoustic description of the data will, obviously, be inferior to those presented above. We will reevaluate the ZCC segmentation from that experiment in a similar manner.

The additional sets of utterances are drawn from two different task domains with much less restricted vocabularies and grammars and are spoken by two different male American speakers. The Allophone (LAL) sentences are 27 general English sentences designed to contain a wide variety of the commonly occuring allophones of that language. [Sho74b] The Btrain set (BTR) consists of 55 sentences drawn from seven, more restricted grammars and task domains. [BakJK75b] In the following table, the somewhat poorer performance shown for BTR is possibly attributable to the different method of hand

	BTR	LAL	TAP
Segments #			
H	1861	1269	1297
M	2990	2157	2290
Missing			
#	253	95	111
%	13.6	7.5	8.5
Extra			
#	1153	819	854
%	62	66	64
d'	.80	1.03	1.01

Figure 6.6: Segmentation Performance -- Other Speakers and Tasks

segmentation used to produce the reference segmentation. In this case, a variant of the Dragon speech understander [BakJK75c] was used to fit standard lexical transcriptions to the signal in a sense defined by the Dragon model to be optimal. These were hand corrected to some extent, but by a different person than the transcriber of the rest of the data. It has not been possible at this point to correct this discrepancy. However, the experience we have had on a wide range of data sets is of the robustness of the segmentation procedures. The excellent agreement between LAL and TAP is found in spite of the fact that thresholds for the segmenter were derived from another corpus, spoken by the speaker of TAP, but the utterances of LAL.

Finally in Figure 6.7, comparison can be made between the first set of results and previously reported segmentation performance as given in Baker and Dixon and Silverman. [BakJM75, Dix75a]

A note concerning interpretation: A careful inspection of Baker's results showed that secondary boundaries increase the total number of hand segments to about 370. We have not generally provided that detailed a hand transcription and, thus, may be reporting as extra boundaries some legitimate detections of secondary segments. Secondly, Silverman and Dixon do not report the sources of knowledge used to produce the segmentation results we have quoted. It is our impression that some amount of label and phonetic rule information is used to improve the segmentation.

	S&D			Baker			SPG
# Segments	6175	216	216	216	150	216	1082
Missing							
#	425	20	40	90	22	56	37
%	6.9	9	19	42	14	26	3.7
Extra							
#	650	38	43	9	51	32	299
%	10.5	18	20	4	34	15	27.6
Shifted							
#	293	--	--	--	--	--	28
%	4.7						2.8
d'	2.73	2.26	1.72	1.95	1.78	1.68	2.38

Figure 6.7: Segmentation Performance -- Other Programs

6.4 Discussion

Although the difficulty of really accurate comparison between different segmentation programs would seem to preclude drawing firm conclusions from the last set of results concerning their relative merits, it is fairly clear that progress is being made. Certainly some programs may be better at detecting and locating certain types of boundaries. At this point, only careful comparison of system errors can provide those insights. However, the major result we wish to put forward is concerned with the parametric representation dimension. There is, indeed, a measurable improvement as one goes to more informationally complete representations of the signal. However, that improvement may not be so critical to system performance as to justify increased computation or hardware† costs. If higher level knowledge can effectively cope with 2 or 3 times the number of extra segments, then we could keep the number of missing segments constant and go from an LPC/DFT computation for SPG to the six analog filters of ZCC. At the ZCC value for d' (1.29), a missing segment rate of 4% corresponds to an extra segment rate of 68% (versus 28% for SPG). Whether or not he is willing to make such trade-offs is entirely up to the system designer.

† Computationally expensive but straightforward representations may be handled rapidly with the use of special purpose hardware at a loss of generality and an increase in initial costs.

It has been our experience in working with the segmentation algorithms developed at C-MU, that, although improvements in performance have been, and will certainly continue to be made, this "snapshot" of the parametric representation dimension is valid for some time to come. We have yet to see promise of any completely parametric level solution to the segmentation problem. This should be obvious upon considering the highly variable nature of continuous speech and the variety of kinds of phenomena to be dealt with in segmenting any single utterance. Thus LIP (Baker) parameters may be effective in locating short burst types of segments while LPC models of the resonant structure may be superior for long, voiced segments and sonorant/sonorant boundaries.

In the final analysis, inspection of specific cases gives the kind of qualitative insights that are also needed to predict performance in a total system. Particular kinds of errors may be very costly if they lead down wrong paths in the search, but only detailed understanding of particular systems can identify such cases. In Appendix S1, we have tried to identify some of the different situations, by presenting cases of discrepancy with the hand segmentation. Where the hand segmentation is correct, other sources of knowledge must be able to override the segmenter's mistake. Where the machine segment best fits the acoustic signal, higher level knowledge must understand such cases of variation from expectation.

6.5 Summary

In this chapter, we have described some of the problems encountered in making a fair evaluation of segmentation. The major problem is acquiring a hand transcription of the correct level of representation. Our approach is to use two referent segmentations at both the acoustic and phonetic level. A model of signal detection, derived from existing theory, gives a useful measure of detectability which normalizes for missing-extra trade-off. The resultant evaluations show a clear preference for high information representations. The SPG and ACS parameters contain a complete model of the resonant structure of the vocal tract impulse response. The ASA and ZCC are approximations to

this information. The first two contain more than 500 bits of information per window, the latter two, less than 100. Finally, we have compared the results obtained with our very low-level segmentation routine to other results reported in the literature. We feel this routine performs quite satisfactorily. With the improvements of phonetic and phonological knowledge used by other programs, it should perform at least as well as them.

Chapter 7

Labeling Procedures

In this chapter, we discuss the labeling procedure adopted for this research. An important issue to be faced is the choice of segmenting or labeling first, and whether the two processes will interact. The choice of distance metrics is also a primary one for this evaluative study. Finally, we introduce some simple parameters which relate to the segmentation; these are prosodic in nature. Training the labeler is discussed, and an algorithm for clustering the training data into natural, intra-phone groupings is presented.

7.1 Role of Labeler, Interface with Segmenter

We have observed that labeling and segmenting are often strongly interconnected in many systems. Indeed, the two processes seem to be two sides of the same coin. Similar techniques are often used to match input patterns with stored templates as are used to match inputs from neighboring time intervals for boundary detection. The choice of whether to segment or label first is, to a large extent, an arbitrary one, often based more upon system structures than upon the requirements of acoustic analysis. However, since we have tried to separate the two processes as much as is realistically possible for the comparative analyses made, it has seemed more sensible to segment first. By associating a number of input patterns with each other in a single segment, this strategy allows one to reduce the sheer number of input patterns which one must compare against a set of stored templates. Labeling second also allows one to make use of the segmentation decisions to locate regions of least acoustic change in the input signal. It is not our purpose to argue the merits of any particular approach to structuring the application of these two processes; it is assumed that segmentation proceeds directly upon the acoustic input. Labeling occurs at those regions of the input which segmentation has selected as being relatively stationary (e.g., pieces of vowels) or as being primitive acoustic gestures (e.g., bursts). The recomposition of these segments, with the labels which they will

acquire, into higher level constructs such as stop consonants, diphthongs, or even phonemes that, although ideally stationary, were realized with internal acoustic variations is a task for a different set of algorithms which utilize phonetic and phonological knowledge. The labeling is intended to be acoustic-phonetic -- it attempts to identify in the input those acoustic patterns which occur during the realization of the phonetic elements.

A second aspect of the labeling process has to do with how one measures correct performance. It is generally considered important for speech understanding systems, that the correct label be "available" among a few alternative choices. Rather than considering only the first choice, most systems will use other sources of knowledge to choose among a few labels for each segment. Thus we are not concerned that the labeler may label a segment with the wrong label, provided it also reports the correct label as an alternative. This kind of requirement has implications mainly for the training process which will be discussed below. A similar way of stating it is to look at the allophonic variation problem. Very often allophones of one phoneme are acoustically very similar to those of another phoneme. Although these separate allophones represent the same acoustic state, we might wish to keep separate descriptions of each -- thus guaranteeing that the labeler will report the "correct" as well as other phonetic labels.

For both reasons, therefore, we have chosen to learn and keep as recognition "templates" the acoustic patterns of a number of variants of each phonetic label. The method of acquisition of the template set and the patterns will be discussed when we present the clustering method below.

7.2 Choice of Metrics

In the survey of pattern classification techniques, a number of distance metrics were described, and the obvious and central role of the distance concept was discussed. Distance in the pattern space occupies just such a central role not only in the pattern recognition but also in the template training methods. Thus, we will briefly restate some of the features of the set of metrics chosen for these experiments.

Euclidean distance (EUC) and Correlation (COR) are each functions of two pattern vectors. Euclidean distance serves to draw spherical loci of equal distance around any point in the sphere. The loci of equal distance around a point for Correlations are cones with the vertex at the origin. Correlation may also be thought of as Euclidean distance in the two dimensional space of the surface of the sphere around the origin. The decision boundaries drawn between any two template points are hyperplanes (perpendicular bisectors of the connecting segment for EUC, through the origin for COR). Although one could consider Euclidean distance to be much more powerful in capturing relationships in the pattern space, Correlations do serve to absolutely normalize out any scalar terms (such as amplitude from a set of filter band parameters).

When second moment data about the templates is available, such as variance of the parameters within each phonetic label class, or covariance (overall or label-specific), more complex distance metrics give somewhat improved results at the cost of more computation. Standard Deviation weighted Euclidean distance (SIG) normalizes each term of the Euclidean distance by the variance in that dimension. Its loci are ellipsoids with axes parallel to the dimensional axes. Finally, if covariance information is available for each label class, we can use the quadratic form to draw boundaries of general quadratic surfaces. This is the Maximum Likelihood metric (LIK) which assumes general Gaussian distributions of the classes and assigns the input to the class most likely to have produced it.

Two other distance metrics have been chosen. The Itakura (ITK) metric is based upon the linear prediction theory and is the estimate of the least squares error term when one interval is predicted by the LPC model of another. The motivation for this type of measure is different from the previously described geometric partitionings of a pattern space; however, it is included in the investigations in so far as it can be applied within the algorithms. Finally, the Baker log probability estimate (BAK) is an ad hoc estimate, based upon the Euclidean distance in a normalized version of the ZCC parameter space, which is of particular interest because of the relatively good results which have been obtained in

the Dragon system using it (even though the ZCC parameters are fairly crude by current standards). [BakJK75b, Low76]

7.3 Prosodic Features

The labeler has available to it the segmentation decisions, and could use these to some benefit if segmental information is known about the training data. Thus, an additional aspect of pattern matching or distance computation which has been useful is a comparison of what may be called the prosodics of the entire segment -- as opposed to the acoustics of the center. We have chosen three rather simple parameters which measure the broad nature of each segment:

1) The average amplitude of the signal over the segment;

2) The duration of the segment;

3) The contour of the amplitude as it compares to neighboring segments.

Calculation of the first two is obvious; the last is merely given a value 1, 0, or -1 if the segment represents a peak, intermediate level, or trough in average amplitude, respectively. A particularly conservative application of this information is made in comparing two segments. A scalar multiplier for the regular distance metric score is composed of the product of three values derived from these three parameters: 1) the ratio of the two average amplitudes, 2) the ratio of the two durations, and 3) the difference between the two contours. These are applied to a (*ad hoc*) function giving values betwen 1 and 2. (See figure 7.1) Thus the minimum scalar is 1 and the maximum is 8. The distance between input and template will be increased when any of these parameters disagree.

While the effect of such prosodic matching is not completely determined when comparing segments produced by different segmenters (hand versus machine), it has proven valuable in clustering, where different duration, contour, or stress may be corollary to allophonic variation.

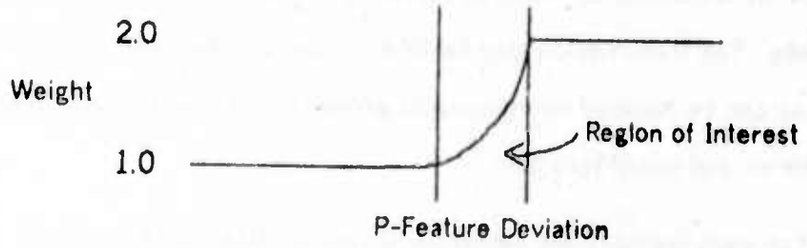


Figure 7.1: Prosodic Weighting Function

In summary, a segment is labeled by computing the distance between its center pattern and each of a set of template patterns, sometimes using the prosodic scalar to penalize the scores of very different "shaped" segments. The closest templates indicate the best alternative labels for the segment. Training uses some of the same metrics.

7.4 Training, Cluster Acquisition

In keeping with the philosophy expressed above, we have designed a training procedure for label templates which discovers the inherent clustering of the sample patterns in the parameter space. The purpose is to identify the acoustic patterns which commonly occur during the realization of each phone.

Since our only model of acoustics is similarity in the pattern space of a particular parametric representation, the algorithm for clustering is based on pattern space distance (and the previously mentioned segment oriented prosodic scalar).

A particular corpus of data has been chosen as the source of samples for training. We have employed for this purpose a set of 27 continuous speech utterances (about 2 minutes) designed to include examples of most common allophones of American English phonemes in semantically and syntactically correct, yet unusual (and thus care-invoking) sentences. These training sentences are recorded under similar conditions for each speaker tested. The approximately 1700 phonetic segments in this corpus have been hand segmented and labeled to an extreme degree of care and fineness of view. This segmentation bears a very strong acoustic flavor in that clear acoustic cues, such as

changes in amplitude or formant shifts, are taken to indicate separate sub-phonemic segments. The transcription may be taken to be as close to an acoustic description of the signal as can be justified by reasonable phonetic theory and careful listening and study of wave forms and spectrograms.

For each segment, the center 10 m.secs. of data yield a vector of parameters in the chosen representations, and the segment boundaries and overall amplitude parameter yield the three prosodic parameters. Each of 40 phonetic labels is represented by an exclusive subset of these 1700 input samples. For each subset, all samples are read in and the complete set of pairwise distances and prosodic scalars is computed. The resultant matrix of distance (x scalar) values is then reordered in the following fashion:

First a threshold is chosen by computing the mean and standard deviation of all the entries in the matrix and setting the threshold to equal MEAN + C x STANDEV (where C is usually -1/2). This data-determined threshold is then applied to each row of the matrix and the row with the most elements within the threshold is selected as indicating the first template. (Ties are resolved by the least sum of all below-threshold elements.) Those samples within the threshold distance of this identified sample are removed from the set. Then we iterate, producing a second, third, etc., template for the particular label, until the entire set is exhausted. (See Figure 7.2 for the resultant matrix of pairwise distances for 89 samples of schwa.) The number of samples supporting each template (the number of remaining samples within threshold) is used to discard "errors" or unusual realizations by discarding templates supported by less than k samples (usually 2, 3, or 4 in these investigations).

Some interesting results have come out of applying this simple algorithm In addition to the particular sets of templates. 1) Errors in the hand segmentation and labeling were clearly identified as poorly supported (often by no other sample) samples of the (errorful) label. 2) While some clearly allophonic distinctions were made, such as de-voicing in /r/ (as in "crude"), often clusters had greater correlations with such factors as position within a word (i.e., stress). We consider all variants of a label which occur frequently enough to

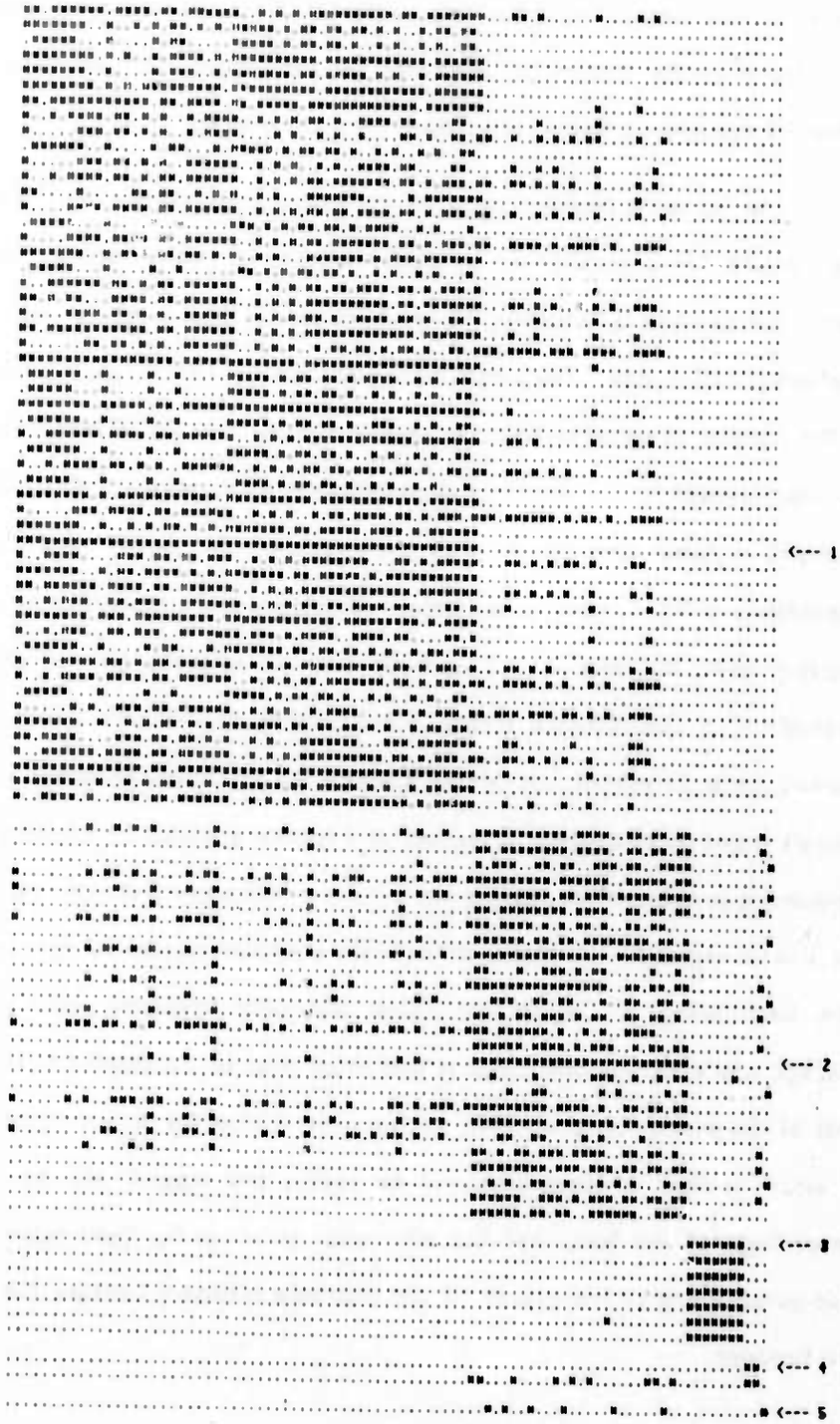


Figure 7.2: Clustering Matrix -- Sorted by Cluster

be useful sources of templates. 3) Finally, even though considerably different parametric representations and distance metrics were used in the clustering algorithm, very small variation was discovered in the total number of templates, and a fair degree of consistency was found in the number of templates for each label. This seems to indicate that the inherent acoustic variations are indeed being discovered.

The following chapter discusses the evaluation of labeling performance and presents the results for a number of parametrization/metric combinations. The most important performance issue is accuracy. However, some attention is due to requirements of storage and computation time. The single most important number to these aspects of performance is the number of parameters in the representation. Storage of templates is linear with this number (except for covariance data which rises as its square). Computation of EUC, COR, and SIG is linear, and of LIK is as the square of the number of parameters. Of similar importance is the number of templates. All storage and computation requirements increase linearly with this number. Speed-ups may be effected by methods such as partial evaluation of the distance metric and discarding of a choice if the partial evaluations exceed some threshold. Or, if the distance to one vowel template is too far, no other vowels might be evaluated. A number of ways are available of speeding up the very time-consuming process of evaluating the full set of distances from the input to each template for every segment. The trade-off between a smaller number of templates yielding a less fine partitioning of the pattern space, and more templates with consequently higher storage and computational costs is one which must be considered in the light of what the rest of the speech understanding system expects from the acoustic-phonetic level analysis. A small number of templates will be costly, and indeed, will be "correct" a higher percentage of the time. Yet the information provided by these fewer templates will be less constraining to the search for the utterance precisely because the classes recognized are broader.

7.5 Summary

The simple pattern classification model is chosen, and a set of basic distance metric are apllied to it. The distances are augmented by a prosodic weight function, composed of three simple measurements of stress and duration of each segment. The less in agreement these parameters are, between the template and the unknown pattern, the greater the distance value will be. Training of the labeler ncludes discovering the inherent clusters within each phone class. This process is, again, based upon the distance metric and pattern space measure of acoustic similarity central to this level.

Chapter 8

Labeling Performance

In order to properly evaluate the labeling performance results, we must first consider some of the issues associated with the labeling process as we have defined it. The following section deals with the problems of deciding 1) what are the objects we are to recognize, 2) what is error and correctness in an ambiguous situation, and 3) what is the effect of segmentation performance on the definition of labeling correctness. Section 2 defines the types of statistics we will present and the experimental dimensions we will cover.

8.1 Some Issues for Evaluation

8.1.1 Recognition Targets

In Section 1 of Chapter 3, we introduced the need to have acoustic-phonetic elements as recognition targets. We decided that, although recognizing phonetic features was also a valid approach to the problem of constructing a phonetic description of the signal, using individual target sounds (hopefully with associated phonetic information) would be more likely to provide the robustness needed by continuous speech understanding systems. In Chapter 7, we presented a method of cluster analysis which could derive these acoustic-phonetic labels as acoustic clusters and representative templates with hand supplied phonetic labels.

A second issue raised at that time pertains to the size of the set of recognition targets -- the fineness of the partitioning of the phonetic space. While we do not deny the existence of such entities as phonemes within the domain of higher level speech knowledge, we have reached the conclusion, along with many others [SchR75, Sil75, Erm74b, Red75b], that the acoustic-phonetic elements found in continuous speech belong to an entire spectrum of such partitionings. That is, to any degree of fineness, there will

always be some ambiguity encountered in labeling (and segmenting) the speech signal, yet, at any degree of fineness, a valid description of the signal can be given. Thus, in choosing where to fix our goals for identification of sounds, we must make some aroitrary decisions. Those decisions may, however, be guided by what we know about the input requirements of knowledge sources at higher levels and about common practice in describing speech at this low level.

The problem of evaluating labeling performance depends upon a set of decisions regarding 1) the samples of the signal used to test -- whose segmentation to employ, 2) the recognition targets -- how fine a description of the acoustic-phonetic states of the signal to create, and 3) the criteria for correctness. This last decision must be made in terms of the system which will eventually use the labels. Thus some discussion of the sources and types of error in labeling is in order if we are to justify how we present the performance results and how we define correct behavior.

8.1.2 Errors

To declare a label in error, we must have at our disposal the "correct" interpretation of the acoustic signal for that segment. Since such information is usually provided by hand segmentation and transcription, we are faced with the problems of level of description and of human error similar to those discussed in Chapter 6 for hand segmentations.

It is clear, for example, that we should not expect the output of even a totally "correct" labeler to match the phonemic content of the utterance, even ignoring discrepancies introduced by acoustic rather than phonemic segmentation to define the location of labeling activity. Vowels will be affected by context -- rounding caused by velerization, centralization caused by lack of stress. Transitions from /z/ to /s/ caused by gradual loss of voicing in final /z/ is another commonly occuring discrepancy.

If there were a clearly accepted set of phonetic rules to explain such deviations from the standard phonemic expectation (and assuming phonological variation was also

handled, perhaps integrated into the lexicon), an automatic evaluation could be used with confidence that the errors found were truly errors of the acoustic-phonetic level.

We are forced, rather, to rely upon a careful hand transcription, where the segmenter/labeler has taken some effort to provide an acoustically valid description of the signal. In spite of its inherent errors and misrepresentations, such a transcription is the best representation of the acoustic-phonetic situation that is currently available. We can only try to make up for incorrect reference labels by providing a few degrees of fineness in the sets of targets for which performance is reported.

8.1.3 Segmentation

We have already raised the problem of segmentation and its effect upon labeling. While we cannot have a perfect segmentation for reference, any more than a perfect labeling, we can improve the correspondence of our reference segmentation with acoustic reality by using a finer, corrected, hand segmentation. In many cases, acoustic segments (with phonemic labels at times) will direct the machine labeler to the relevant portions of the signal. Sonorant segments will be composed of more or less steady state sub-segments, which are the places where we can expect the best labeling performance. Their locations are not always available at the phonemic or phonetic level.

The procedure we have adopted is to acquire the best hand segmentation possible at the lowest level of description possible. Then a corrected segmentation, such as the one refered to in Chapter 6 and Appendix S2, is merged with this transcription†. The "correct" label for these segments is considered to be the hand label which was in effect at the middle of the acoustic segment.

Labeling experiments are, therefore, performed over the set of segments (and thus samples in the pattern space) which we would most like our segmentation routines to

† If such a correct segmentation is not available, it is better to use the machine segmentation of the best parametric representation, with its segmentation errors, but tuned to miss as few segments as possible, than to use a broad, hand segmentation and label over transition portions of the signal.

provide in the actual system. We will additionally show some results of labeling the segments provided by the machine segmenter, although we will not evaluate these completely. It is hoped that the construction of a labeling reference which combines our expectations for acoustic segments and phonetic labels will most accurately reflect the expectations or needs of a speech understanding system at the higher levels of analysis.

8.2 Evaluation Space

The design space which we are attempting to investigate is quite large. Even after we have accepted the limitations discussed in Chapter 1 -- the choice of four parametric representations, ignoring the multiple speaker normalization issue, restricting ourselves to acoustic pattern space labeling, and the choice of simple distance metrics with static training procedures -- we are faced with the issues of recognition target set size, error criterion, and segmentation just discussed. Instead of spending any more time justifying the decisions we have made -- it is probably sufficient to have raised the issues -- we will outline the dimensions to be covered and the methods of presentation of results.

8.2.1 Experimental Dimensions

The four parametric representations will be used for a set of 40 sentences from the News Retrieval task (also used for the segmentation evaluation). In addition, a second set of News Retrieval sentences, spoken by another male American speaker, and third set, used in the development and testing of the Dragon system [BakJK75b], will be evaluated.

With the first set, as many of the basic distance metrics -- EUC, COR, SIG, and LIK -- as are applicable will be used over the SPG, ASA, and ZCC parameters. The modified ZCC parameters used by Dragon will also provide a point of reference for one total system's performance. The ACS representation is to be used with Itakura's specially designed log ratio measure only, as it is poorly suited for the more standard distance functions. (See Ichakawa, [Ich73] for the poor performance of Partial Correlation Coefficients.)

The purpose of the additional sets of utterances is, of course, to justify our

assertion that the training and labeling process as a whole is valid for more than one speaker and set of utterances. However, the additional data can only help to improve the fidelity of our performance results.

Finally, for each experiment, we will present a family of results for a sequence of phone class sets (see Figure 8.1). Each set provides a partitioning of the phonetic space at a different degree of fineness, and may be relevant to particular aspects of the total speech understanding problem.

8.2.2 Methods for Presenting Labeling Accuracy

Schwartz and Makhoul [SchR75] point out that the appropriate response to the problem of ambiguity in continuous speech is ambiguity in the segmentation and labeling output. That is, optional interpretations of the input signal should be put forth as possible alternative recognitions rather than one and only one label for each segment and one stream of segments.† If the labeler finds a number of plausible labels for a single speech sample, it can do no better than to rate and return them all. Thus, it is not sufficient to evaluate labeling accuracy without indicating the criteria for accepting a label as such a plausible interpretation of an input pattern. Indeed, returning a number of labels may merely be considered to represent a finer partitioning of the pattern space. There is a duality, which is similar to the feature vs. template issue, between returning alternative labels and using more and finer ones.

In lieu of any more global information, the labeler can only use the pattern space distance value with which it orders the templates to rate them as well, and to select the

† While we do not provide a lattice structure of segments and labels as in the Speechlis system, neither do we claim to make use of phonetic rules or sub-phonemic segment sequences. The acoustic segments are detected with a strong bias towards the extra segment end of the ROC curve, and the next level in Hearsay II, for example, is able to combine many of them in an optional manner within the flexible data base structure of Hearsay II. Within this optional segment structure, sets of alternative labels are re-combined according to a phonetic feature calculus to produce a new set of most likely labels for the combined segments.

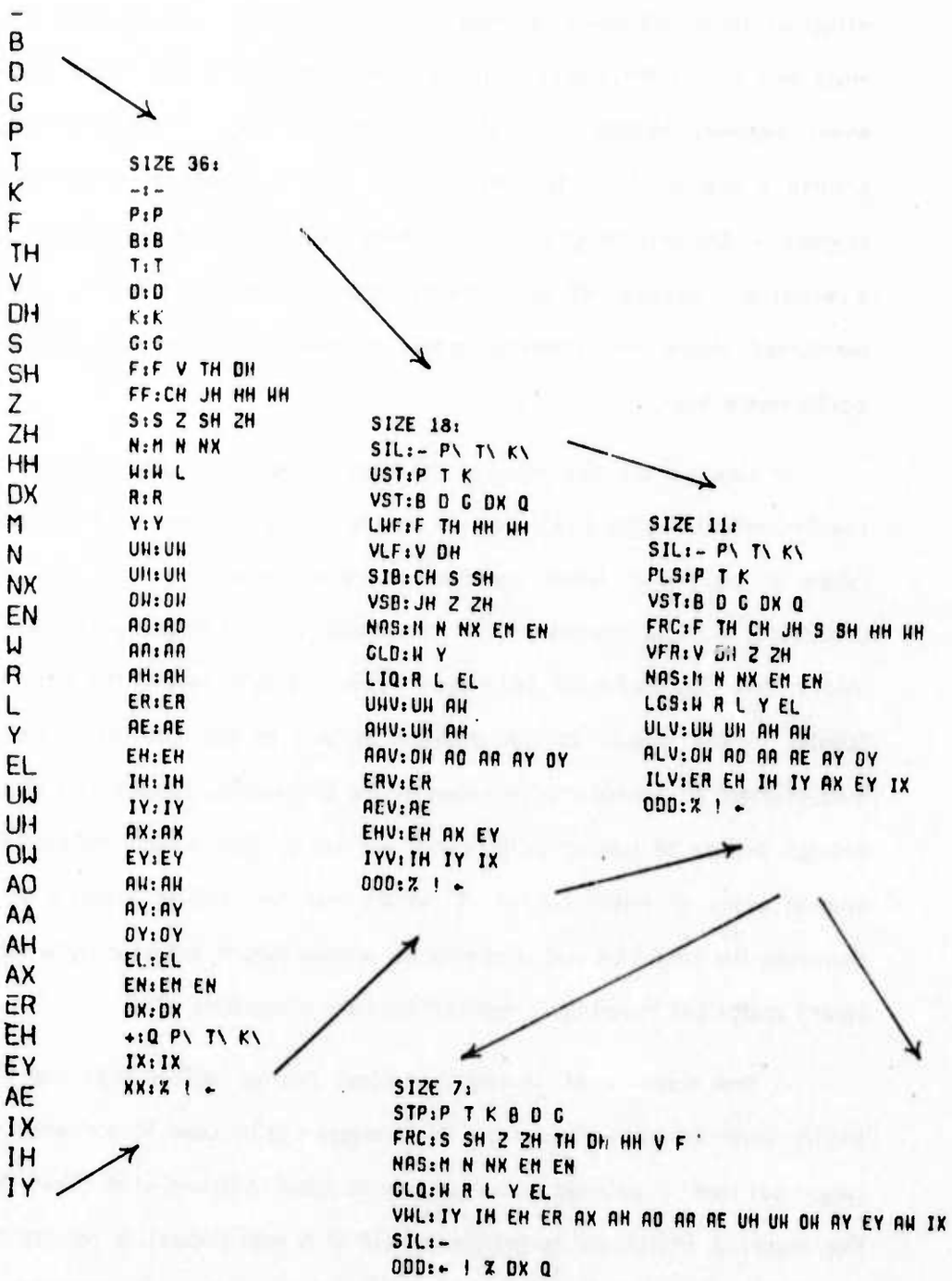


Figure 8.1: Phone Class Sets

acceptable ones.‡ If we define a selection criterion (e.g. $d(x,y) \leq \min_y (d(x,y)) + T$) then we can collect the statistic: Pr{correct label passes criterion}. In order to understand the effect of the acceptance of multiple labels on system performance, size of search, etc., we must also collect the statistic: E[number of recognition objects which pass criterion] for every segment labeled. This latter may be considered as a measure of the factor of growth a search of the possible optional machine transcriptions displays at each new segment -- the branching factor (BF). These statistics can be presented in graphic form as a relationship between BF and Accuracy. It ought to be noted that the recognition objects mentioned above are whatever labels or classes of labels we choose to evaluate performance over.

In case we are interested in the kind of errors made, a confusion matrix provides the Pr{recognize y|input x}, where x and y are both members of the particular set of labels or classes of labels under consideration. However, there is some difficulty in presenting multiple choices in this format without producing a great deal of extraneous information. The confusion matrix does, however, give an idea of the error behavior of the labeler -- the quality of the mistakes as well as the quantity. Combined with the measurement of correctness provided by the BF/accuracy graphs, this should be a broad enough picture of labeler performance and yet provide enough detailed information for special cases of interest. The BF values may be used to estimate system resource demands; the confusion matrix conditional probabilities to estimate higher level confusions (word confusions based upon incorrect phonetic information, etc.).

A final measure of labeling proficiency can be derived from the signal detection theory referred to in Chapter 6. This measure can be used to normalize for recognition target set size. A detailed discussion may be found in Green, et al. [Gre64], but, basically, the accuracy, Pr{correct target chosen out of N possibilities}, is related to the same d' measure presented earlier -- the difference between the signal and noise distributions in

‡ We do, in fact, use certain global information in the form of prosodic features mentioned in Chapter 7. The effect of these features is, however, integrated into the pattern distance value as soon as the templates are matched with the unknown input pattern.

some signal measure space. The d' values given below are computed from tables provided in [Gre64] and based upon a theoretical model which assumes all the N targets to be orthogonal in the pattern space. Since this is not true, because of our need for phonetic templates which may duplicate one another to some degree, the accuracy rates are consequently lower, and the d' values lower, than predicted by the model for the actual signal to noise ratio. They can, however, serve as an interesting normalized measure for comparison of labeling performance, just as they do for segmentation performance.

8.3 Results of Labeling -- One Speaker

There are so many dimensions to even a simple labeling experiment that we will explore the space along each one, individually, rather than try to cover the entire set of possible labelers. This section presents the results of labeling experiments performed on the 1416 segments which comprise the 40 news retrieval sentences dealt with in Chapter 6. The dimensions of interest are: Parametric Representation, Distance Metric, Acceptance Criterion (Branching Factor), and Target Set Size. In addition, confusion matrices and hand analysis are presented to provide a qualitative picture of some typical labeling performance.

Figure 8.2 shows overall labeling accuracies for the four parametric representations.

p	SPG	ASA	ZCC	ACS
1	24.6(1.0)	27.1(1.0)	20.3(1.0)	28.7(1.0)
2	42.4(1.9)	39.1(1.9)	31.4(1.9)	44.4(1.9)
3	54.0(2.8)	50.4(2.8)	42.0(2.8)	54.6(2.7)

Figure 8.2: Labeling Performance -- Different Parametric Representations

The distance metric is the Euclidean distance function†, and the full set of 40 recognition targets is evaluated. The values reported for position p are Pr{correct template in position ≤ p}. The expected number of different targets (branching factor) is given in parentheses. In these results, concerned with speaker CC, the clustering algorithm of Chapter 7 provided 63 templates for SPG data, 76 for ASA, 75 for ZCC, and 87 for ACS. All are acoustic templates for the 40 phonetic targets.

Note that, although ASA is a bit better than SPG for p=1, SPG improves faster with increasing position (BF). This may be due to one or two bad templates for SPG which "capture" first place often enough to affect that statistic. More careful tuning of the template sets, while desirable, could not be done for all experiments. The ZCC representation is the only clearly inferior one, much as might be expected from its few, broad filters. Yet its performance is not too much worse than the others.

Figure 8.3 shows the representation fixed at ASA, and presents four distancegure metrics. Again, the accuracies are given for the first three positions. The same template set was used throughout.

p	EUC	COR	SIG	LIK
1	27.1(1.0)	25.0(1.0)	28.7(1.0)	25.1(1.0)
2	39.1(1.9)	37.0(1.9)	41.3(1.9)	35.5(1.6)
3	50.4(2.8)	49.3(2.8)	50.6(2.8)	44.1(2.3)

Figure 8.3: Labeling Performance -- Distance Metrics

Almost identical performance is obtained from EUC, COR, and SIG. The LIK metric makes use of more information (the covariances within each target training set), yet is unstable for some targets. This is due to instability of the covariance matrix inversion

† The ACS representation was run only with Itakura's log ratio distance.

caused by insufficient samples in the training set. It should be noted, however, that the BF values for LIK are lower at each position, indicating a greater likelihood of multiply-recognized targets. For the same BF values, LIK performs comparably.

In both of the above dimensions, there is very little difference among the choices examined. We will discuss this apparent lack of preference in the last section. It is a rather strong result of this work.

Figure 8.4 is a graphic display of accuracy versus Branching Factor for the SPG/SIG experiment.

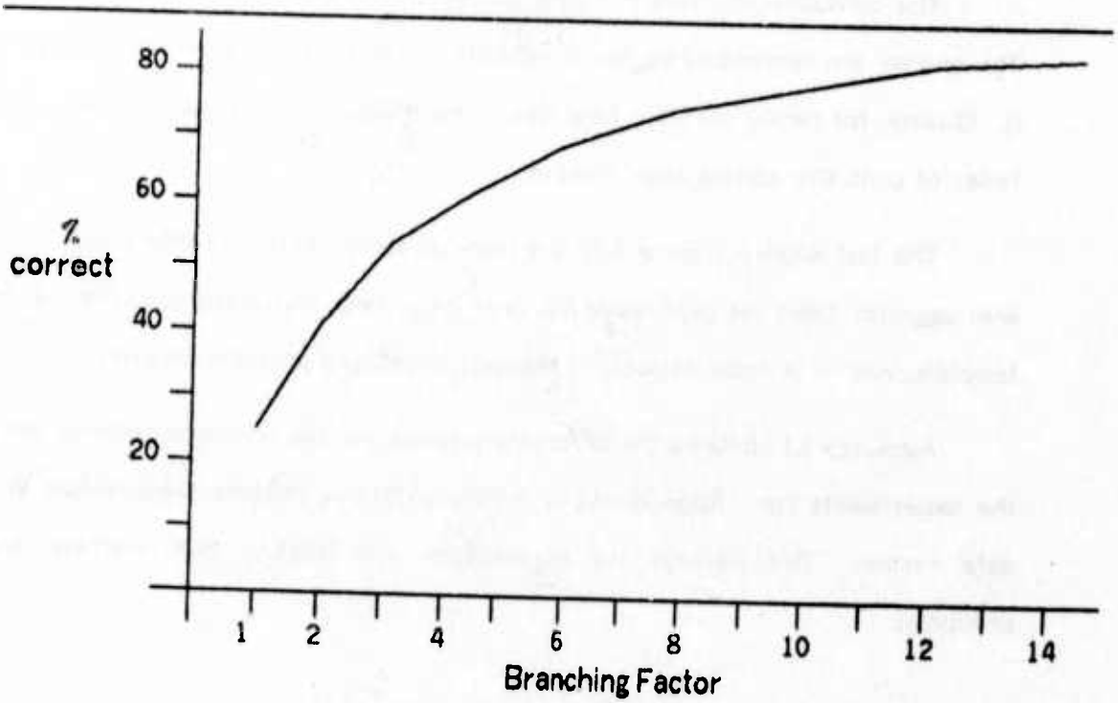


Figure 8.4: Branching Factor vs. Accuracy

Five plots are given, identified by the size of the target set used in each evaluation. The BF plot gives a particularly convenient view of accuracy against the demands that will be made upon higher levels by excess options in recognition.

A normalization may be made for target set size by the signal detection model discussed earlier in this chapter. The figures for BF=1.0 from the SPG/SIG experiment are

Size	40	29	17	10	6
Pr (C)	25.6	31.8	36.6	43.0	60.3
d'	1.4	1.5	1.4	1.3	1.6

Figure 8.5: Effect of Size of Target Set

given in Figure 8.5 with their theoretical detectability, d'†.

The confusion matrices in Figure 8.6 were obtained from the SPG/SIG experiment. The entries are normalized by row to estimate, for entry ij, Pr{target j in position 1 | hand l}. Clearly, for rarely occuring hand labels, this estimate will be less accurate. Important types of confusion can be seen, however.

The last display (Figure 8.7) is a trace of a representative utterance. The entries are: segment times (in centi-seconds), hand label, rank and score (distance) of "correct" template, and -- in order of score -- template, score, and prosodic weight*10.

Appendix L1 contains the BF/accuracy tables and the confusion matrices for each of the experiments run. Appendix L2 is a more extensive machine transcription of the CC data corpus. This displays the segmentation and labeling, both machine and hand produced.

8.4 Results of Labeling -- Other Speakers and Vocabularies

In the last section, we extensively explored the evaluation space for one speaker and task (vocabulary). In order to extend the validity of our results accross both the speaker and vocabulary dimensions, we have run a limited set of additional labeling experiments.

These include a second speaker (LE), again for the news retrieval task and

† Calculated by approximation -- the procedure may be less accurate for small size. (see [Gre64])

	-	B	D	G	P	T	K	F	TH	V	DH	S	SH	Z	ZH	HH	DX	M	N	NX	EN	W	R	L	Y	EL	UW	UH	OW	AO	AA	AX	ER	EH	EY	AE	IX	IH	IY
-	70	7	2	5	5	.	1	1	.	1	1	.	.	.	.	.	.	2	1	.	.	.	.	.	3	.	.	.	.	.	.	1	.	.	.	.	.	1
B	12	36	10	2	14	.	.	.	.	.	12	.	.	.	.	.	.	12	.	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.
D	9	14	16	4	12	5	.	.	.	4	.	.	.	.	.	2	4	5	2	.	.	.	2	.	14	.	.	.	.	.	.	2	.	.	.	4	.	2	2
G	5	9	21	.	7	5	14	.	.	2	.	.	7	.	5	.	2	9	2	.	.	.	5	.	7	.	.	.	.	.	.	.	.	.	.	.	.	.	.
P	25	.	.	6	44	.	.	.	.	6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	13	.	.	.	.	.	.	6	.	.	.	.	.	.	.
T	6	2	4	.	16	20	6	4	.	4	.	.	10	.	2	6	4	.	.	.	.	.	.	.	4	.	.	.	.	.	.	.	2	.	2	.	.	.
K	.	12	.	.	3	.	33	3	.	.	.	.	27	3	3	3	6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	5	.	.	.	.	.	.
F	37	.	5	11	16	.	.	11	.	.	.	.	5	.	.	5	.	11	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TH	10	25	.	10	25	.	.	10	5	.	.	.	.	.	.	15	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
V	3	.	3	.	.	.	.	.	.	3	3	.	.	.	.	.	20	13	3	.	.	13	.	.	30	.	.	.	.	.	.	.	.	9	.	.	3	.
DH	14	14	29	7	7	.	.	.	.	.	7	.	.	.	.	.	.	7	7	.	.	.	.	.	.	.	.	.	.	.	.	7	.	.	.	.	.	.
S	.	2	.	.	6	2	22	2	.	.	.	31	10	9	.	2	2	.	.	.	.	.	.	.	5	.	.	.	.	.	.	.	.	.	.	.	2	.
SH	.	.	.	.	.	.	67	.	.	.	.	33	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Z	.	3	3	.	.	3	10	.	.	3	.	29	12	26	.	.	.	.	.	.	.	.	.	.	3	.	.	.	.	.	.	.	.	.	.	.	.	.
ZH	.	.	.	.	.	.	25	.	.	.	.	.	50	.	.	.	.	.	.	.	.	.	.	.	25	.	.	.	.	.	.	.	.	.	.	.	.	.	.
HH	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	33	17	.	.	.	.	.	.	.	33	.	.	.	.	.	.	.	.	.	.	17	.	.	.
DX	.	.	10	.	.	.	.	.	.	.	.	.	.	.	.	.	40	.	.	.	.	.	.	.	40	.	.	.	.	.	.	.	.	10	.	.	.	.	.
M	4	.	14	.	.	.	.	.	.	.	14	.	.	.	.	.	4	36	11	10	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
N	.	1	37	.	.	.	1	.	.	.	.	.	.	.	.	.	7	15	13	23	.	.	.	.	1	.	.	.	.	.	.	1	.	.	.	.	.	.	.
NX	.	.	11	.	.	.	.	.	.	.	.	.	.	.	.	.	.	5	.	74	.	.	.	.	.	.	.	.	.	.	.	5	.	.	.	5	.	.	.
EN	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
W	.	.	13	.	.	.	.	.	.	.	.	.	.	.	.	.	.	7	7	.	67	.	.	.	7	.	.	.	.	.	.	.	.	.	.	.	.	.	.
P	.	.	.	.	2	.	.	.	.	.	.	.	2	.	.	.	2	4	4	2	.	11	13	.	2	.	2	.	.	2	2	.	44	2	.	.	2	.
L	2	2	.	4	2	.	.	.	.	.	.	.	.	.	.	.	2	.	4	6	.	7	2	26	.	11	.	.	.	6	2	11	.	6	2	2	4	2	.
Y	.	.	.	.	.	.	.	.	.	.	.	.	13	.	.	.	.	6	.	.	.	.	.	.	38	.	.	.	.	.	.	.	.	.	.	19	.	13	13
EL	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	100	.	.	.	.	.	.	.
UW	3	.	.	.	.	.	.	.	.	.	.	.	.	3	.	.	.	.	.	3	.	.	.	.	10	5	13	3	.	.	.	3	.	.	0	33	.	13	5
UH	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	.	.	.	.	.	.	0	.	.	15	39	15	.	.	.	15	.
OW	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	0	.	23	.	.	0	23	.	15	.	.	.	.	.	15	.
AO	.	.	.	.	.	.	.	.	.	3	.	.	.	.	.	.	3	.	.	.	23	7	17	.	17	.	.	3	.	7	7	10	.	.	3	.	.
AA	.	4	.	4	.	.	.	.	.	4	.	.	.	.	.	.	.	.	.	.	.	4	.	.	0	.	0	.	0	0	38	.	.	.	0	.	0	.
AX	1	1	2	.	2	.	1	.	.	2	.	.	.	1	.	.	1	2	.	1	.	.	1	3	2	1	0	1	9	3	30	2	2	.	7	2	16	2
EP	.	.	.	4	4	.	.	.	.	13	.	.	.	.	.	.	4	.	0	.	.	.	4	.	.	.	4	.	.	.	.	33	.	.	4	4	0	0
EH	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	2	.	7	.	.	.	1	9	5	1	15	.	31	.	25	1
EY	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	20	.	.	.	.	.	.	.	.	.	80	.	.	.
AE	.	.	3	.	3	.	6	.	.	3	.	.	.	.	.	.	6	.	.	.	.	.	6	.	3	.	.	.	3	.	3	6	6	.	44	.	0	3
IX	.	.	6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	6	6	.	.	.	.	25	.	.	.	.	.	25	6	.	.	6	6	.	13
IH	.	.	1	.	.	.	1	.	.	.	.	.	.	.	.	.	1	.	.	1	.	.	.	.	25	.	.	.	1	.	1	1	6	.	38	.	17	5
IY	.	1	1	2	1	.	.	.	.	.	.	.	.	.	.	1	1	1	.	.	.	1	.	.	26	.	.	.	1	.	6	.	1	.	15	.	26	16

40 x 40

	SIL	UST	VST	LWF	VLF	S1B	VSB	NAS	GLD	L1Q	UWV	AHV	AAV	ERV	AEV	EHV	1YV	DDD
SIL	70	6	13	1	2	.	.	3	3	.	.	.	.	.	.	1	1	.
UST	7	40	12	7	3	18	3	.	4	.	.	.	.	.	1	4	.	.
VST	8	18	39	1	6	2	1	9	10	2	.	.	.	.	1	2	1	.
LWF	20	18	24	24	.	2	.	4	4	.	.	.	.	.	2	.	.	.
VLF	7	2	32	.	7	.	.	16	30	.	.	.	.	.	.	5	2	.
S1B	.	31	3	3	.	49	9	.	4	.	.	.	.	.	.	.	1	.
VSB	.	21	5	.	3	42	24	.	5	.	.	.	.	.	.	.	.	.
NAS	1	1	35	.	3	.	.	58	1	.	.	.	.	.	.	1	1	.
GLD	.	.	6	.	.	6	.	10	52	3	.	.	.	.	10	.	13	.
L1Q	1	2	5	.	.	1	.	10	10	27	1	.	6	20	1	12	4	.
UWV	3	.	.	.	.	.	3	3	10	5	13	3	.	.	33	10	10	.
AHV	.	.	.	.	.	.	.	0	.	.	0	15	15	.	38	15	.	.
AAV	.	.	4	.	3	.	.	.	10	29	.	3	13	3	3	25	7	.
ERV	.	4	0	.	13	.	.	0	.	4	4	.	.	33	4	.	21	.
AEV	.	0	0	.	3	.	.	.	3	6	.	3	6	44	0	11	.	.
EHV	.	2	3	.	1	.	.	1	5	2	.	5	11	2	10	27	22	.
1YV	.	1	4	1	.	.	.	2	26	.	.	.	1	1	25	0	31	.
DDD	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

18 x 18

	STP	FRC	NAS	GLO	VWL	S1L	DDD
STP	53	19	6	7	5	0	3
FRC	30	42	5	10	3	6	4
NAS	31	3	58	1	2	1	6
GLO	5	2	10	41	39	1	2
VWL	4	2	2	17	73	.	1
S1L	19	2	8	3	2	70	.
DDD	10	.	.	40	10	.	40

7 x 7

Figure 8.6: Confusion Matrices, SPG/SIG

"TELL ME EVERYTHING ABOUT ENGLAND"

TIME	PHONE		TEMPLATES RECOGNIZED											
65:67	T	1 21	T1	21 10	K2	28 18	P1	20 12	SH1	31 14	B2	32 15		
67:69	T	2 15	B1	12 12	T1	15 15	TH1	16 10	G1	16 10	P1	17 10		
69:75	EHtLt	3 19	AA1	14 20	IH3	17 20	EH2	19 29	EH1	19 20	AO1	21 20		
75:79	L	3 21	AO1	15 30	AA1	19 20	L2	21 40	ER1	22 40	R1	23 40		
79:86	M	4 24	M2	21 14	NX1	21 19	W1	22 14	M1	24 19	V1	30 14		
86:105	IY	1 25	IY1	25 22	AE3	31 22	IH4	34 29	AE1	37 45	AA1	38 51		
105:108	-	1 17	-2	17 10	B1	18 14	-1	20 10	D1	22 12	TH1	23 11		
108:115	EH	3 22	AE3	16 20	Y1	20 10	EH1	22 20	IH1	22 23	IY1	22 20		
115:118	EH	1 14	EH1	14 40	AE3	20 41	IH1	20 25	ER1	21 40	IH3	21 23		
118:122	V	2 23	DX1	16 15	V1	23 10	W1	23 17	N2	23 12	NX1	26 16		
122:126	R	3 22	ER1	18 23	UW1	18 10	R1	22 10	DX1	23 10	IH2	24 20		
126:128	IY	7 43	IH1	29 36	AX1	35 40	IH3	38 23	EH1	38 30	L2	41 30		
128:133	IY	2 25	Y1	23 10	IY1	25 20	IH4	20 20	AE3	31 20	IH1	37 20		
133:142	TH	8 24	B1	15 10	HH1	17 10	-2	17 10	-1	19 10	D1	20 10		
142:158	IHtNt	6 24	AE1	15 41	EH1	16 73	AE3	18 40	IY1	19 22	AA1	23 50		
158:167	NX	1 11	NX1	11 25	N1	17 17	M1	17 14	N2	17 25	W1	21 24		
167:174	AXt-Nt	1 15	AX1	15 23	K1	16 12	IX1	17 21	IY2	19 20	AX3	21 20		
174:178	-	1 11	-2	11 10	B1	14 10	-1	15 10	G1	19 17	M2	22 20		
178:185	AX	2 15	IH2	14 38	AX1	15 25	ER1	19 29	IY2	20 20	EH1	21 20		
185:191	B	5 15	D1	12 10	DH1	13 11	-2	14 10	M2	14 20	B1	15 10		
191:206	AW	-----	AO1	19 32	AA1	22 53	AE2	23 42	IH2	23 44	AE1	24 31		
206:208	T	19 53	DX1	19 15	D1	20 10	V1	29 12	R2	31 20	DX2	31 16		
208:211	I	-----	IY1	25 20	AE3	25 33	IH4	25 20	ZH1	26 10	IH1	27 20		
211:214	I	-----	B1	26 10	HH1	27 10	F1	30 10	D1	31 11	Y1	32 24		
214:220	IHtNt	2 18	Y1	18 10	IH4	18 21	IY1	22 20	ZH1	31 10	K1	32 10		
220:224	IHtNt	1 34	IH4	34 40	IY1	37 40	AE3	40 21	AX1	40 39	AE1	41 40		
224:230	NX	1 15	NX1	15 10	M1	21 10	N2	22 19	N1	23 10	G1	24 10		
230:232	G	2 14	D1	12 11	G1	14 10	DH1	16 15	N1	18 25	M1	20 35		
232:239	L	38 --	NX1	14 12	N1	16 10	G1	17 10	D1	19 10	M1	19 11		
239:244	AXtNt	3 25	P1	23 11	T1	24 10	AX3	25 20	K1	26 11	IX1	27 20		
244:259	N	27 --	M2	10 40	-2	11 11	B1	14 15	-1	15 10	EN1	53 20		
259:282	DTHt	10 --	G1	19 19	TH1	19 23	P1	19 18	B1	33 14	F1	30 22		

|	|	|	|	|Prosodic Weight * 10
|	|	|	|Distance Score	
|	|	|Template		
|	|Distance to Correct			
|Rank of First Correct Template				

Figure 8.7: Trace of Labeling Evaluation

vocabulary, and a third speaker (JB) with partly different task influences. The JB sentences were run by the Dragon system and were all recognized correctly using the BAK distance metric -- a modified Euclidean distance. Most important to that performance was the carefully tuned word lexicon, which provided a great deal of phonological disambiguation.

Figure 8.8 gives the results for recognition of the full phone set, in the first three positions, for these data sets, as well as the ZCC/EUC results reported above for speaker CC.

Speaker	CC	LE	JB	JB	JB
Task	AP	AP	FRM&AP	FRM	FRM
Parameters	ZCC	ZCC	ZCC	ZCC	ASA
Metric	EUC	EUC	EUC	BAK	EUC
# Segments	1416	932	1144	732	732
p=1	20.3(1.0)	21.8(1.0)	40.7(1.0)	42.8(1.0)	39.9(1.0)
2	31.4(1.9)	36.9(1.9)	53.5(1.8)	53.8(1.7)	48.0(1.6)
3	42.0(2.8)	44.0(2.8)	57.0(2.6)	62.6(2.4)	56.6(2.4)

Figure 8.8: Labeling Performance -- Other Speakers and Tasks

There are a few observations to be made. First, the performance of the LE data is almost identical to the CC data. This is in spite of the fact that many more (120) templates were found for the LE training†. However, the conditions under which the recordings were made and the hand segmentation and labeling performed were the same for both CC and LE data. Thus, the level of representation of the expected labels was very similar for both data sets.

In the case of the JB data, the hand referents were actually generated by a modified form of the Dragon system. This form sought to fit the correct sentence to the

† A totally arbitrary variation, due to a different cluster rejection criterion.

signal (using the ZCC/BAK decision metric) by applying all the knowledge that Dragon would have at its disposal for recognition. Then obvious errors were hand corrected. Clearly, this expectation is closer to the ZCC/BAK results. However, it is also closer to the other acoustic, machine results tested here than a totally hand produced transcription would be. The fact that the level of representation of the evaluation referent, and not the decision metric and parametric representation, is the cause of this increased accuracy is shown by the last column. The same data run on ASA/EUC, neither used by the machine transcription form of Dragon, produced the same high performance results as the ZCC/EUC experiment.

8.5 Discussion

A short discussion of the preceding labeling evaluations is in order. The lack of significant differences among a large number of the experiments might seem rather counterintuitive in the light of the differences among the parametric representations in segmenting. However, the tasks of segmentation and labeling are quite different, and we believe this lack of comparative difference to be an important result. In addition, the labeling may seem to be performing at a rather low level in comparison to other systems such as were discussed in chapter 4. We claim that this level of performance is, in fact, reasonably good performance for the current state of the art. It merely needs to be evaluated at a lower level of representation than has been done.

In a preliminary study of labeling accuracy, we found considerable differences among both representations and metrics. The experimental set-up was, however, quite different. One template per target sound was acquired by averaging a number of training samples. Testing was then performed over the same data set. In such a situation, second moment data greatly aided the SIG and LIK metrics in identifying the training populations correctly. In addition, averaging had a more disastrous effect upon some parametric representations than others.

In the current experiments a significant amount of knowledge about the distribution

of patterns for speech sounds is contained in the multiple templates for each target sound. Thus, the theoretical shape of the decision boundaries is of much less significance in affecting error rates. Since we do not test and train on the same set, performance is poorer (more realistic), and the specific distribution of the training samples is less significant.

Errors of a higher level are probably responsible for the performance shown in the last section. Coarticulation, effects of phonological and phonetic variations, prosodic effects, and other sources of confusion are all beyond the capacity of a simple template matching routine. It is significant, in fact, that the labeling performances are so much alike. This similarity indicates that most of the action available to acoustic level labeling is being achieved.

A second issue is the low accuracy reported in the last section. The explanation is clearly the lack of any higher level knowledge in our labeling routine. But to support our claim to reasonable performance of this simplified labeler, we can point out the following. First, the Itakura log ratio metric has been tested in a word recognizer by Itakura [Ita75] and has yielded excellent results for limited speech recognition tasks. The same parametric representations and metric yield less than 30% accuracy at the acoustic level, and close to 98% at the word level.

A second point is Baker's Dragon System.[BakJK75b] The parametric representation and distance metric used are essentially ZCC/EUC (with some amplitude normalization). This classification function was tested and gave results comparable to the results reported above. Templates generated by our clustering routine were substituted for the standard phonetic spellings used by Baker in his initial development. In addition, phonological rules were applied to the lexical entries for mistaken words to produce alternative template sequences for those words. However, none of this tuning was performed on the test data. Dragon was run on a set of 578 words in 102 sentences from five tasks, one speaker, with a dictionary of 354 words. The word level accuracy reported was greater than 99%. Phonological and syntactic knowledge sources were sufficient to correct all the errors of a 30% accurate labeler.

When the Dragon system knowledge sources are used (in a modified form of the system) to generate a transcription of the known utterance, they produce a referent transcription more suited to evaluating acoustic labeling. Comparison with this referent yields accuracies of 40% and greater for the least capable parameters, ZCC.

Finally, we may refer to Shockey and Reddy's foreign language transcription experiment [Sho74a] and to Klatt and Stevens' spectrogram reading results [Kla72]. Our results are quite comparable with trained phoneticians reading spectrograms or waveforms in the absence of any higher level linguistic support. They are not much worse than human performance with auditory input.

Chapter 9

Conclusions

In this concluding chapter, we would like to draw together some of the previously discussed results and methods into a more coherent view of speech recognition activities at the parametric level. At the same time, we should also restate the major contributions of this work and point out particular areas where future effort is warranted.

We will first offer a brief summary of the entire thesis, in order to draw together some of the major points and restate the primary results. We will then list the contributions with short discussions, and finally proceed to discuss the parametric level of speech understanding systems in the light of this work. The last section will be devoted to possible areas for further research.

9.1 Summary of the Thesis

This thesis is a study of machine speech recognition at the parametric level. It attempts to evaluate and understand the relative merits of a number of alternative design choices at that level. Such a study raises issues in Artificial Intelligence, Linguistics, Acoustics, Pattern Recognition, Statistics, and Speech Understanding research. In particular, it involves an investigation of segmentation and labeling techniques, and the use of parametric representations for the acoustic signal in those techniques. Every speech recognition system employs some parametric representation and some initial signal-to-symbol transformation. We show the performance currently available for these initial processes, and assert that such performance is comparable to human performance. We present the relative merits of some typical parametric representations, and develop a methodology for such comparative evaluation. Simple, parameter-independent schemes for segmenting, labeling, and training are developed as well. The role of pattern classification techniques, as they relate to the initial signal-to-symbol transformation, is clarified.

9.1.1 Background

Although most of our knowledge about how to recognize and understand speech is taken from human performance, the structure of computer speech recognition and understanding systems is of particular importance to this study. Knowledge about speech is generally organized into separate sources of knowledge; each works with a representation of the information content of the input utterance. These representations may exist at a number of different levels, as suggested by their elements: speech sounds, phonetic gestures, phonemes, syllables, words, syntactic units, concepts, etc. In evaluating the performance of recognition processes at the parametric representation level, we eliminate, as much as possible, the effects of ambiguities from other levels. Such ambiguities as coarticulation or phonological variation will strongly affect the degree to which the expected transcription of an utterance corresponds with the acoustic performance. A great difficulty in comparing published results is that the level of the knowledge used in recognition and the representation used for evaluation are not usually specified. Usually, only total system performance may be compared, not the effectiveness of component methods.

9.1.2 Parametric Representations

Parametric representations fall into a few major types; typical examples of each have been chosen for study. A bank of broad-band filters (ZCC) with amplitude and zero-crossing measurements, and a bank of narrow-band filters (ASA), amplitude only, represent analog methods. A digital Fourier transform of the LPC filter [Mar72] produces a smoothed spectral envelope (SPG) very much in current use. Finally, the autocorrelation sequence (ACS) is employed with a special method designed for it.[Ita75] Each method yields a set of measurements at uniform, short intervals -- a pattern.

9.1.3 Distance Metrics

Distance functions, chosen from Pattern Classification theory, are then applied to the parameter patterns as measures of acoustic similarity. The basic model adopted is that of a vector of parametric measurements for each pattern. These vectors define a space of possible patterns; within this space a measure of distance may be applied between patterns. As populations of sample patterns are accumulated, better statistical descriptions may be estimated of the true distribution of those patterns in the space. A simple example might be to collect all the occurences of a phone and compute the mean and variance of each dimension. Then a suitable measure of similarity might be Euclidean distance, weighted by variance, to approximate a measure of the deviation from population mean. This is one distance metric chosen (SIG). The others are Euclidean distance (EUC), Correlation (COR) -- the magnitude normalized dot product, and Maximum Likelihood (LIK). In this last, the population covariance matrix is used to calculate Pr{unknown produced from population}, under the assumption of Gaussian distributions.

9.1.4 Segmentation

A method for segmenting speech into isolated, acoustically consistent segments is presented. The method is fairly independent of the choice of parametric representation, since it relies upon the acoustic similarity measure as the primary evidence of acoustic change. First, however, a threshold is applied to the signal amplitude measurement to discriminate between speech and silence. Then the speech portion is examined further. In collecting evidence for a segment boundary, a measure of change is applied to neighboring parameter patterns. This measure produces a time sequence of values whose peaks are detected and subjected to a threshold for acceptance or rejection. A composite of such functions yields the final segmentation. Narrow and broad pattern similarity and amplitude change are the three functions applied to the non-silence portions of the signal. This process is very much like the process hypothesized in the basic model for Signal Detection

[Ega64]. That model may be applied to the problem of evaluating segment boundary "detectability."

Missing and extra segment errors are found to be as good as 4% and 19%, respectively. Significant differences in the segmentation effectiveness of the parametric representations is found. They may be ordered as follows: SPG, ACS, ASA, and ZCC. The best performance is found to be comparable to the state of the art. Little reduction in accuracy is encountered when new speakers are tested.

Figure 9.1 shows the results of segmentation for 40 sentences from the News Retrieval task, one speaker.

	SPG	ACS	ASA	ZCC
Segments #				
H1	1082	1082	1082	1082
H2	1541	1541	1541	1541
M	2026	2082	2010	2290
Missing				
#	37	57	91	52
%	3.7	5.8	9.2	5.3
Extra				
#	299	391	434	681
%	27.6	36.1	40.0	63.0
	(19.4)	(25.3)	(28.1)	(44.2)
Shifted				
#	28	34	45	41
%	2.8	3.4	4.5	4.1
d'	2.38	1.93	1.58	1.29
	(2.65)	(2.24)	(1.91)	(1.77)

Figure 9.1: Segmentation -- Different Parametric Representations

The reference segmentation contains 1082 segments, primarily at the phonemic level of description. The second reference contains corrections to this file (1541 segments), to make it more an acoustic description of the corpus. The number of machine segments

reported may be greater than the sum of this second size (hand reported acoustic segments) and the number of extra boundaries. The discrepancy is an artifact of the way we evaluate segmentation. Occasionally, two machine boundaries will fall close enough to a hand boundary so that both are accepted. Such segments must, therefore, be very short, and are usually transition segments which may easily be deleted at higher levels. The number of missing boundaries (segments), divided by the number of boundaries which are included in both reference segmentations, is the missing segment error rate. The number of shifted boundaries is also divided by this number. The number of extra boundaries is divided by the number of primary segments (the size of H1 in this case). The extra segment rates in parentheses are those where division is by the number of acoustic segments (size of H2). The value, d', is a single measure of detectability from the Signal Detection model. It has the effect of normalizing for the trade-off between missing and extra segment errors.

9.1.5 Labeling

Labeling is accomplished by simple pattern distance metrics. Given a set of phonetic elements as the recognition targets, a set of templates for each target is derived from the training data. This is achieved by a clustering algorithm developed for the purpose of encoding into the set of templates some of the ambiguities encountered because of allophonic variation. The pairwise distances are computed for all pairs of sample patterns in the training population for a particular phonetic target. Then a threshold is chosen from these values, and the distances below threshold are marked. The sample pattern in the most marked pairs is chosen as a representative template and all its marked mates are discarded. After iterating, the population is divided into clusters of various sizes, each with a "best" representative template pattern. Clusters of sufficiently small size are ignored.

Labeling itself proceeds by computing the distance from the unknown pattern to each template. In addition to the distance metrics mentioned, three prosodic features of each segment -- the average amplitude, the duration, and the amplitude contour of the

surrounding segments -- are used tc increase the distances to templates whose prosodic features are considerably different.

The set of templates (and their appropriate target labels) and the distance scores give the total recognition information available from this straightforward labeler. If some criterion is placed on the templates which one is willing to report to the rest of a system, then accuracy may be measured as a function of the severity or looseness of that criterion. If the true effect, upon a speech recognition system, of loosening the acceptance criterion is to be understood, one must also measure the expected number of separate targets reported at each instance. We call this the Branching Factor (BF), and collect it as well as accuracy statistics in evaluating labeling performance.

Little difference is observed along the parametric representation or the classification metric dimensions, except for poorer performance for ZCC input. Each input segment is labeled as one of a set of 40 phone labels. The correct phone appears as the first choice 28% of the time. It appears in the first three choices 55% of the time. However, when a lower level, acoustic transcription is used as the evaluation referent, these values increase to 42% and 65%. Even the 28% accuracy, which arises from a comparison against phonemic expectation, is acceptable performance; it is the same as or slightly better than human spectrogram reading performance in the absence of other linguistic clues.[Sho74a]

Figure 9.2 shows overall labeling accuracies for the four parametric representations. The distance metric is the Euclidean distance function†, and a set of 40 phonetic recognition targets is used. The values reported for position p are Pr{correct template in position ≤ p}. The expected number of different targets (branching factor) is given in parenthesis.

Figure 9.3 is a graphic display of accuracy versus Branching Factor for the SPG/SIG experiment. Five plots are given, identified by the size of the target set used in each evaluation. The BF plot gives a particularly convenient view of accuracy versus the demands that will be made upon higher levels by excess options in recognition.

† The ACS representation was used only with Itakura's log ratio distance.

p	SPG	ASA	ZCC	ACS
1	24.6(1.0)	27.1(1.0)	20.3(1.0)	28.7(1.0)
2	42.4(1.9)	39.1(1.9)	31.4(1.9)	44.4(1.9)
3	54.0(2.8)	50.4(2.8)	42.0(2.8)	54.6(2.7)

Figure 9.2: Labeling -- Different Parametric Representations

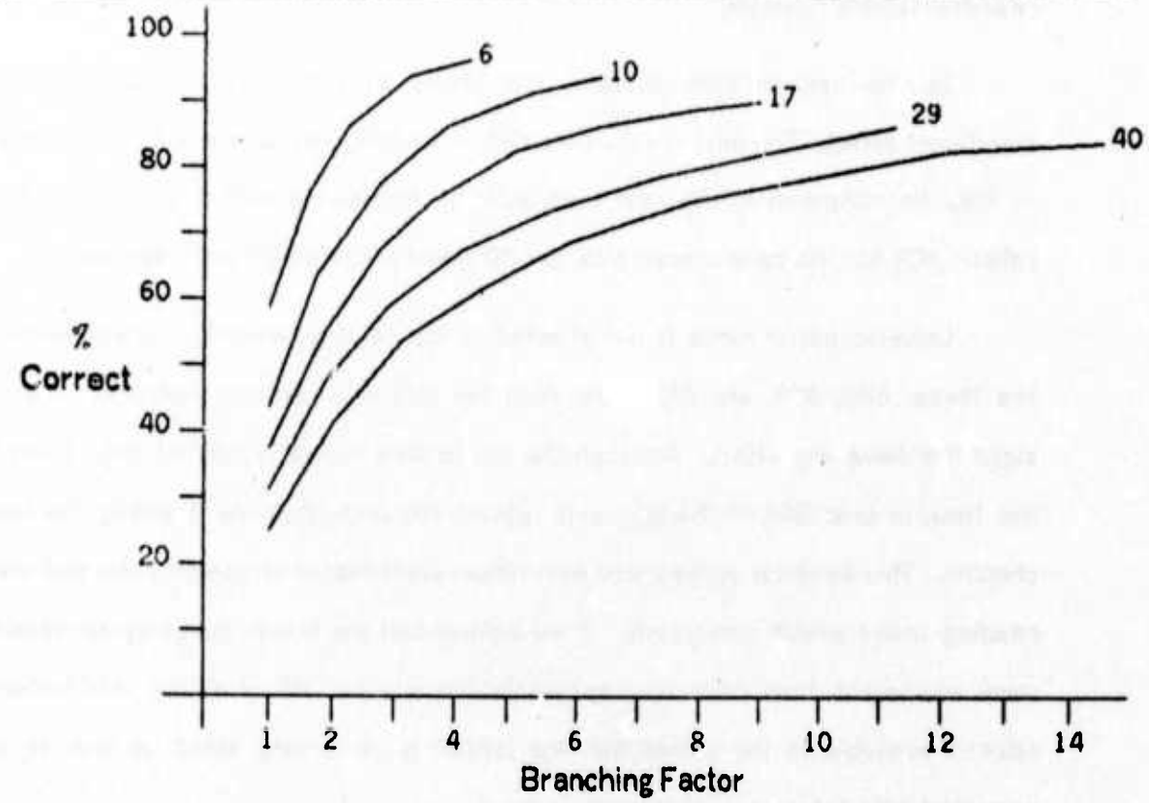


Figure 9.3: Branching Factor vs. Labeling Accuracy for Various Target Sets

9.2 Contributions

9.2.1 A Comparison of Parametric Representations

It should be clear that any effort to compare the various parametric representations in terms of their suitability for use in speech recognition systems is needed, and such

results are of value. What is not clear is how accurately such comparisons have been made and with what confidence they can be applied to predicting performance and designing systems. It is our contention that the uniform manner in which we have applied the tested representations to segmentation and labeling, the simplicity and ubiquitous nature of the pattern classification assumptions, the quantities of data, and the care with which we have evaluated the results reported in this dissertation all contribute to the fidelity of those results and to our confidence in, at least, the relative strengths of the representations reported.

In the segmentation process, the choice of parametric representation shows significant effect. The best results obtained -- for SPG, missed rate = 4%, extra rate = 19% -- may be compared to the next best, ACS, by normalizing with d' for one of the error rates. ACS for the same missed rate as SPG would produce 33% extra segments.

Labeling performance is not affected by choice of parametric representation among the three: SPG, ACS, and ASA. Nor does the choice of distance metric for the labeling algorithm have any effect. Although the top labeling choice is correct only about 25% of the time, in over 50% of the segments labeled, the correct choice is among the top three choices. This behavior agrees well with human performance at spectrogram and waveform reading under similar constraints. If we believe that the human spectrogram reading was very competent, then while the representations may not afford all the information about speech available to the human ear, the labeler is performing about as well as may be expected with those representations as input.

9.2.2 Parameter-Independent Segmentation

The segmentation algorithm described in Chapter 5 may be easily adapted to any set of parameters and any measure of similarity in the space of those parameter vectors. The algorithm need not be trained for every speaker, and works well for a variety of parametric representations. Our experience with it has indicated very little degradation of performance accross speakers, with recording conditions constant. This is probably due to

the fact that boundary detection depends upon detecting changes in the pattern space. Large, sudden changes will be detected by any reasonable scheme for segmentation. The small or slow changes are the most difficult to detect correctly. Yet, these small shifts are not affected by large, smooth transformations to the overall pattern space which may characterize speaker change. (e.g. Different format locations, for a new speaker, will not seriously affect the detection of format shifts.) Labeling, on the other hand, depends upon gross comparisons of patterns to a much greater extent; it is more dependent upon the absolute locations in the space of patterns for particular phone templates. The segmenter represents an available tool and a benchmark for acoustic level segmenting whose overall performance is comparable to other current programs. Moreover, the method of threshold acquisition is easily adapted to more dynamically sensitive techniques, as will be mentioned below.

9.2.3 The Role of Primitive Pattern Classification Methods

In the past, many of the methods and results available from statistical pattern classification research have been dismissed or tacitly assumed of small value without sufficient attempt to understand the implications of the surrounding issues (training, target sets, metrics, etc.) which strongly affect performance. It is hoped that this work will stimulate further, careful application of the methods involving stochastic pattern spaces. It is apparent, however, that much of the disenchantment with these techniques stems from their failure to solve the recognition problem at a level accessible to higher level knowledge sources. Pure statistical classification approaches will not, in our opinion, provide such a solution, as our low initial labeling accuracies indicate -- 25% may sound low to the uninitiated. The analysis of the proper roles for pattern classification methods which we have presented, in the context of testing their usefulness, may also serve to define fruitful avenues for applying more sophisticated pattern classification techniques. A great deal has been written about the acoustic level of speech, but fairly little attention has been paid to techniques which are specifically and specially suited to computer implementation. We feel that pattern classification methods are so suited. One interesting

result of our work, with direct implications for pattern classification techniques and speech, is the irrelevence of second moment statistics for describing the template distributions -- the distance metric dimension. For example, ASA parameters were labeled with the four metrics: EUC, COR, SIG, and LIK, the last two utilizing variance and covariance statistics, respectively. When accuracy results are calculated for a fixed branching factor of 3.0, all accuracies are within 1% of 50%. Either the clusters are fine enough divisions of the pattern space to capture all the relevent information about the target population distribution without the need to involve second moment data, or, alternatively, the distributions are spherical (i.e. the parameters are uncorrelated). It is hard to imagine adjacent narrow filter bands in the range of frequencies in question to be uncorrelated, so it is likely that the former argument is more valid. This implies that the emphasis for speech applications of pattern classification techniques should be in clustering, tracking, or dynamic training -- to capture, empirically, the complexities caused by stress, coarticulative, and other phonetic variations.

9.2.4 Methodology for Evaluation

Closely related to our view of the role of pattern classification is the methodology we have adopted for evaluating performance in those pattern spaces. We have essentially found that, in order to evaluate accuracy fairly, one must have a fair representation of what is expected or correct. As an example, labeling accuracy over a set of 40 phones increases from 20% to 40% when the labeling referent is supplied by a machine aided process rather than purely hand transcription. The machine aided process find the best fit to the acoustic input of label templates, constrained by the stored phonological variations in a word lexicon. Although this method yields higher accuracy measurements, we have not used it because it is open to question. The referent is being generated primarily by the same process which is to be tested. However, we believe that the machine aided referents generated are as valid descriptions as are the hand transcriptions. Much of our difficulty has been in acquiring hand segmentations which represent a fair expectation for acoustic level performance.

The methods and attitudes presented are applicable to any parametric representation, to any segmentation and labeling output, and to a number of levels of description of speech. Where knowledge in addition to acoustic/phonetic knowledge is used, the results will be more like higher level representations and may, therefore, yield higher accuracies when compared to hand transcriptions. For the IBM labeling and segmentation program, reported accuracy at the phonemic level is 62%. The program uses digital spectral parameters and a pattern matching scheme very similar to our own as input to a detailed set of phonetic and phonological rules. Such performance results are valid measurements to be applied to the analysis of total system performance, even though they may imply less about particular aspects (such as a particular parametric representation) than our primitive level evaluations.

9.2.5 Signal Detection Model

Applying the model of signal versus noise, and the d' measure of signal detectability [Tan64] may prove quite useful for modeling errors of raw segmentation and labeling output over a wide range of performance trade-offs. Notably, extra versus missing segment errors, and recognition set size can be normalized by d', and whole ranges of performance predicted. While the validity of this model has not been completely established, our preliminary success with it, added to the large amount of human perception research supporting it as a model of detection, seem to lend it credence. We also see possible applications in predicting performance of systems under simulated errorful inputs, prior to implementing the actual knowledge sources.

9.2.6 Clustering

Our success with the acoustic/phonetic clustering algorithm gives us hope of even further gains to made in this direction. By using multiple templates for various acoustic manifestations of phones, we are able to describe complex partitionings of the pattern space with simple metrics. In addition, we are able to factor out effects which alter the acoustic, but not phonetic, nature of the signal. Dynamic methods for tracking clusters may

be applicable here. At minimum, this routine provides a method for integrating two levels of representation, acoustic and phonetic, which are often difficult to correlate. Once again, the routine is applicable to any <pattern space, distance metric> definition of similarity of two speech samples.

9.3 Parametric Representations

The design or choice of a set of acoustic parameters for speech recognition analysis is still a difficult problem. We have shown how accuracy is improved -- usually at the cost of increased computation and memory requirements -- by choosing more informationally complete representations. However, this is not the only source of computational costs. Lower accuracy, in most systems, introduces larger data bases and more extensive searches at higher levels. Thus a total system analysis of costs in memory and speed requirements versus accuracy must be made by the system designer if the choice of parametric representation is to be made with cost in mind.

At the present state of the art, emphasis has been placed mainly upon accuracy, since that aspect of performance is the most critical to a number of the goals of speech understanding systems, even to overall speed and memory. However, if systems are to be designed for limited resources, low cost, and real-time operation, excessive parametric information must be trimmed away. In another sense as well, parametric information should be as sparse as is necessary to meet the system performance goals. This is in order to reduce the likelihood that extraneous aspects of an input pattern will lead to error†. A number of methods for selecting parts of the parametric pattern, according to *a priori* decisions about speech class, are available, from sequential decision methods [Fu68] to specific parameters designed for such classes [Wei75, Ata75].

Of the performance information reported in Chapter 8, the confusion matrices are the most useful to a system designer concerned with special cases -- specific situations

† For example, in a situation where one is reasonably sure of a fricative sound, a lot of information about resonant structure in the lower frequencies is worse than useless. It may actually cause mislabeling to a high vowel if there is any voicing present (as there often is).

where particular parametric representations may fail or succeed. An extensive analysis of sub-matrices such as is found in Weinstein, et.al. [Wei75] should really be structured to match the particular higher level knowledge and particular requirements of processes found in each system. We cannot give specific recommendations of the type, "Use parameters X for case A...," since the cases of interest are determined by the individual systems.

The overall performance results do reflect actual, continuous speech recognition of American English sentences. In that respect, they reflect the *a priori* distributions of phonetic types, coarticulation situations, stress and pitch variations, etc., which are likely to be encountered under similar conditions of speech. The results are, therefore, more valid for prediction than if they had been compiled from artificial word lists or from a smaller data corpus. In the light of such a belief, the results indicate the relative effectiveness with which segmentation and labeling can be performed at the most primitive level of recognition, averaged over a number of different situations. Since most knowledge sources will build upon primitive decision mechanisms, we feel the comparative results reported here will be valuable even for more sophisticated, phonetic level speech recognition programs.

9.4 Parametric Level Knowledge Sources

In comparison with some of reported work at the acoustic/phonetic level, our segmentation and labeling routines may seem rather harshly limited to the parametric representation level only. However, our view has been that other knowledge can be applied by separate processes at separate times if the system structure is sufficiently flexible. This is neither a new, nor an extremely insightful, point of view, but it does allow us to focus on the set of recognition decisions which occur prior to any phonetic or phonological analysis. It also is a logical extension of the concept of modularly implemented separable sources of knowledge so often expressed in the literature. [New71, Red73, Erm74b, Les75, Woo75]

We suggest that a reasonable approach to analysis of speech at this level is to make the transformation from parametric representation to acoustic/phonetic segments as soon as possible and with as much information as is relevent. The machine transcription occuples considerably less space than the complete parametric representation of the signal. In addition, the kind of processing needed to create this transcription is straightforward and easily performed in a parallel manner, perhaps with special purpose machinery, or off-line, in cases where experiments are run before or during system development.

We have shown that, by using simple pattern classification techniques, reasonable labeling and good segmentation performance may be achieved. Using these simple pattern space measures for many decisions yields the additional bonus of parameter independence. The routines are not built to accommodate particular parameters, but rather designed to make use of the information inherent in the occuring populations of entire pattern vectors. Thus, the method of extracting parameters may be changed during system development, or after, whenever better methods are found, and the routines may be expected to work well without extensive re-tuning.

We are not, however, arguing against the use of more complex decision procedures, nor against more feedback from higher level knowledge sources. Rather, we stress the need to make as complete and valid use of the patterns of parameters as is possible, as early (low level) as possible. This requires detailed knowledge of the statistical nature of the pattern space, encoded in the trained templates, distance measures, and related aspects of the pattern classification functions.

9.5 Evaluation

Our approach to parametric level processing by simple pattern classification requires that one disengage higher level prejudices and expectations from one's evaluation of the accuracy of such methods. Our philosophy for performance evaluation has been to expect what is really in the input to appear in the output transcription, and, additionally, to

expect absent what is not in the input. For example, if we consider only individual parameter vectors extracted at a single short interval of time, we cannot hope to integrate into our decisions segment-sequence phonetic information such as is available in stop consonants. This is not to say that such complex patterns are not in the input as a whole, just not in the particular input to the decision rule being evaluated. Rather, we suggest that the right time to determine whether such information is preserved by the lower level routines is when higher level knowledge sources are evaluated. To continue the example, since a /t/ burst is acoustically similar to an /s/, we are somewhat satisfied if our labeler returns /s/ for some of the /t/ bursts. It is the job of the phonetic/phonemic level knowledge sources to discover /-//s/ sequences and label them /t/. Such cases will lead to a lower overall accuracy score for our labeling evaluation, since we do not have such knowledge encoded in our labeling referent. However, the confusion matrix entries of /t/ for /s/ and vice versa should be recognized as less critical by anyone investigating the labeling accuracy of a particular parametric representation.

In view of the previous discussions, our performance measurements would appear to be the lower bounds of performance to be expected from a particular representation. We feel that such a lower bound is as valuable a measure as more optimistic estimates which integrate the results of some higher level knowledge sources. Certainly, if the total systems are to be modeled in terms of individual processes, such a "separation of power" view is necessary.

The idea of modeling the entire system is particularly attractive but difficult to accomplish. As knowledge source interaction has become more complex, our understanding of the implications of errors at various levels has become less complete -- more derived from the special cases actually traced. The signal detection model may provide a (very broad) model of this lowest level of recognition activity -- less detailed than the confusion matrix model of errors, but easier to manipulate. We can foresee the d' detectability measure in use to parametrize a zeroth order simulation of segmenting and labeling. This model might be improved by applying the conditional probabilities available in the

confusion matrix. As similar models are developed for other levels, overall knowledge interaction schemes can be simulated.

9.6 Topics for Further Research

It has become almost obligatory in many dissertations such as this one to include a list of topics for further research. However, we would like to include such a list for quite a different reason than tradition. In the course of these investigations into parametric level processing of speech for machine recognition and understanding, we have been made aware of a number of interesting possibilities for extending or improving techniques for segmenting and labeling as well as some interesting approaches to evaluating performance and the problems of training the parametric routines. We believe that a great deal of progress may be made in these areas, and have had some difficulty in keeping to a particular path of research with all the tempting problems surrounding this level. Moreover, we have spent some effort in presenting a view of the current state of the art, and that view will not be complete without pointers to the aspects most likely to yield further progress.

9.6.1 New Parametric Representations

The search for new and better parametric representations will, of course, continue. Particular models of speech production or reception in humans, such as the all pole LPC model, or Baker's LIP parameters† [BakJM75], will continue to provide new insights into the kind of information and encodings found in human speech. Another direction yields parameters and decision procedures designed to detect specific phonetic features. [Ata75] It is important to consider the decision procedure to be used with a particular parametric representation, for the effective shape of the pattern space depends upon both.

Where does this lead for future parametric representations? Perhaps a more integrated approach to their development will result from such considerations -- one in

† based upon neuropsychological evidence of zero-crossing responsive cells

which the needs of machine recognition of speech, and machine oriented higher level features, play as strong a role as aspects of human perception have played in the past. Certainly, any new parametric representation must be extensively tested and compared with the already numerous available ones if we are to make real progress. In this, the work reported here will serve as a valuable tool for guiding research.

9.6.2 Segmentation

A number of different segmenters are currently being developed and tested, and it is quite likely that features of many will prove particularly useful to others. The ideas expressed in our segmenter for integrating boundary detections from a number of functions of the signal may prove useful for integrating segment evidence from a variety of sources of such knowledge. However, one problem which seems likely to yield to immediate beneficial solution is that of adjusting detection thresholds (or whatever tuning parameters are relevant to the particular segmenter in question) to the non-stationary behavior of speech. Boundaries are characterized by a variety of durations, magnitudes, and qualities of signal change. It may be necessary to extend the period of time over which the signal is viewed, to adjust the thresholds to reject insignificant changes, or to ignore entire regions of the pattern space, if they are independent of the phonetic information in the signal.

A powerful line of attack is suggested by recent work in visual segmentation. [Ohl75] In this approach, histograms of the parametric measurements are analyzed for each scene, in order to determine the most likely parameters for segmentation as well as the best thresholds for those parameters (for that scene).

In speech research, a very low level segmenter has been added to the Dragon system to improve speed of recognition [Low76]. Good success has resulted from a single detection parameter which tracks the change over a varying time period, looking further during slow changes for evidence of acoustic boundaries.

The message here is that a number of statistical pattern classification techniques

seem to be applicable to the segmentation problem. Dynamically adjusted, self-training detection routines will result in very robust, high performance segmentation.

9.6.3 Recognition Targets

In a very similar sense as segmentation, better training of recognition templates will, undoubtedly, result from dynamic tracking of input data in the manner discussed by a number of pattern classification researchers. We wish to point out another aspect of the target problem which needs attention at this time -- the integration of acoustic and higher level knowledge about speech in the recognition target set. This problem has been extensively discussed by others [Wei75] and efforts have been made to construct, *a priori*, sets of phonetic labels which are acoustically distinct. We feel that such sets must be discovered in much the same way as other aspects of the pattern space, by statistical analysis of bodies of data. Clearly, there are many improvements to be made to the simple clustering algorithm of Chapter 7. We look forward to more positive results from data-derived recognition targets, and, in addition, from data-derived higher level rules [Smi75, Hay75]

9.6.4 Evaluation

With regard to evaluation, there is so much to be done that we will limit our discussion to one important aspect of evaluating accuracy performance at the parametric level. That is the problem of acquiring high-fidelity referents to which recognition results may be compared. Since the performance of knowledge sources at one level of speech may not be expected to match the expectations of another level, performance evaluations will be in error unless the level of description (of the signal's contents) of the referent and the recognition are quite similar.

An obvious procedure is to take great care in hand producing the referent transcriptions. Not only is this extremely laborious†, but it also fails because the human

† which leads to errors and limits the quantities of data that may be analyzed

transcriber may not understand just what the expectations and implications are of a knowledge source encoded in rules, programs, or trained statistics. It would appear that the best source of these expectations is the speech understanding system itself. The time has come for systems to be designed and implemented with evaluation of various knowledge sources as a basic facility. At each level of representation, facilities should be available to derive, from the knowledge sources, just what inputs from other sources would result in the correct action, decision, etc. As an example, if a higher level of the system can recognize the transition segment /l/ in lower vowels following /g/ or /k/ as an indication of those stops, such cases in the referent should include /stop//l//vowel/ as an alternative to /g,k//vowel/.

It may well be time to depart from the close ties to human perceptual experience with speech. Some of the most successful systems to date, both for word recognition and connected speech understanding, [Ita75, BakJK75b, Erm74b] have had much less in common with what we know about humans and linguistics than with what we know about computers and artificial intelligence techniques. It is less important to model human processing than to match human competence; especially since we know so little about the elements of the human speech perception mechanism. To this end, the best use must be made of quite different information processing devices than humans seem to have, and of different forms of data and control.

9.7 Envoy

This thesis has involved investigations of a number of design choices for the acoustic/parametric level of computer speech recognition. It has led us to survey a large range of techiques, and to attempt to extract aspects of Pattern Classification, Acoustic Analysis, and Performance Evaluation most relevent to the stated goal -- a comparison of segmentation and labeling performance. The current efforts to develop speech understanding systems are producing in their wakes a number of theories about speech. Although overall performance of a total system is an important measure of the validity of

its assumptions, the difficulty of studying each system's components, *in vitro*, has been a handicap to the entension of our understanding of the entire problem. This work is an attempt to extract one basic component and evaluate it in a manner which will both aid designers and increase understanding of its role in the total speech recognition problem.

References

[Ais64] M. A. Aiserman, E. M. Braverman, and L. I. Rozonoer, "Theoretical Foundations of the Potential Function Method in Pattern Recognition," *Avtomat. i Telemeh.* 25, 917-936, 1964.

[And62] T. W. Anderson and R. Bahadur, "Classification into Two Multivariate Normal Distributions with Different Covariance Matrices," *Ann. Math. Stat.* 33, 422-431, 1962.

[Ata71] B. S. Atal and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," *J. Acoustic Society Amer.*, 50, 2, 637-655, 1971.

[Ata75] B. S."Atal and L. R. Rabiner, "A Pattern Recognition Approach to Voiced Unvoiced Silence Classification with Applications to Speech Recognition," unpublished document, Bell Labs., 1975.

[BakJK73] J. K. Baker, "Machine-Aided Labeling of Connected Speech," in *Working Papers in Speech Recognition II*, Tech. Report, Computer Science Dept., Carnegie-Mellon University, Pittsburgh, Pa., 1973.

[BakJK75a] J. K. Baker, "The DRAGON System -- An Overview,". *IEEE Trans. ASSP*, 23, 24-29, 1975.

[BakJK75b] J. K. Baker, "Stochastic Modeling as a Means of Automatic Speech Recognition," Ph.D. Thesis, Computer Science Department, Carnegie-Mellon University, Pittsburgh, Pa., 1975.

[BakJM75] J. M. Baker, "A New Time-Domain Analysis of Human Speech and Other Complex Waveforms", Ph.D. Thesis, Carnegie-Mellon Univesity, Pittsburgh, Pa., 1975.

[Bau71] F. L. Bauer and C. Reinsch, "Inversion of Positive Definite Matrices by the Gauss-Jordon Method," Contribution I/3 in *Linear Algebra*, J. H. Wilkinson (ed.), Springer-Verlag, New York, 1971.

[Bru68] D. J. Bruce, "Effects of Context upon Intelligibility of Heard Speech," in Oldfield and Marshall (eds.), Pp. 123-131, 1968.

[Chi66] Y. T. Chien and K. S. Fu, "A Modified Sequential Recognition Machine using Time-varying Stopping Boundaries," *IEEE Trans. Information Theory*, 12, No. 2, 206-214, 1966.

[Coc67] W. T. Cochran, *et. al.*, "What is the Fast Fourier Transform?," *Proc. IEEE*, 55, No. 10, Oct. 1967.

[Coo74] F. S. Cooper, "Acoustic CUes in Natural Speech: Their Natures and Potential Uses in Speech Recognition," Research Proposal, Haskins Laboratories, New Haven, April 1974.

[Dix75a] N. R. Dixon and H. F. Silverman, "A General Language-Operated Decision Implementation System (GLODIS): Its Application to Continuous Speech Segmentation," Report RC5368, IBM, T. J. Watson Research Center, Yorktown Heights, N. Y., 1975.

[Dix75b] N. R. Dixon and H. F. Silverman, "Some Encouraging Results for General Purpose Continuous Speech Recognition," *Proc. 1975 Int. Conf. on Cybernetics and Society*, San Francisco, Ca., 293-295, 1975.

[Dru73] H. Drucker and J. Preusse "Real Time Recognition of Ten Vowellike Sounds in Continuous Speech," Ft. Monmouth, N.J. and Monmouth College, N.J., 1973.

[Dud73] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, John Wiley and Sons, N.Y., 1973.

[Ega64] J. P. Egan, A. I. Schulman, and G. Z. Greenberg, "Operating Characteristics Determined by Binary Decisions and Ratings," in Swets (ed.), Pp.172-186, 1964.

[Erm74a] L. D. Erman (ed.), *Contributed Papers of IEEE Symposium on Speech Recognition*, Carnegie-Mellon University, Pittsburgh, Pa., (IEEE cat. no. 74CH0878-9 AE), 1974.

[Erm74b] L. D. Erman, "An Environment and System for Machine Understanding of Connected Speech," Ph.D. Thesis, Computer Science Dept, Stanford University, Technical Report, Computer Science Department, Carnegie-Mellon University, Pittsburgh, Pa., 1974.

[Erm75] L. D. Erman, "Overview of the Hearsay Speech Understanding Research," *Computer Science Research Review*, Computer Science Dept., Carnegie-Mellon University, Pittsburgh, Pa., 1975.

[Fan74] G. H. Fant, unpublished comment, 1974.

[Fei63] E. A. Feigenbaum and J. Feldman, *Computers and Thought*, McGraw Hill, New York, 1963.

[Fla72] J. L. Flanagan, *Speech Analysis, Synthesis and Perception*, Springer Verlag, New York, 1972.

[For74] J. W. Forgie, D. E. Hall, and R. W. Wiesen, "An Overview of the Lincoln Laboratory Speech Recognition System," *J. Acoustic Society Amer.*, 56, S27 (A), 1974.

[Fu68] K. S. Fu, *Sequential Methods in Pattern Recognition and Machine Learning*, Academic Press, N.Y., 1968.

[Gil74] R. A. Gillmann, "Automatic Recognition of Nasal Phonemes," in Erman (ed.), 1974.

[Gre64] D. M. Green, T. G. Birdsall, "The Effect of Vocabulary Size on Articulation Score," in Swets (ed.), 1964.

[GolB59] B. Gold, "Machine Recognition of Hand-sent Morse Code," *IEEE Trans. Information Theory*, Vol. IT-5, Pp. 17-24, March 1959.

[GolH74] H. G. Goldberg, D. R. Reddy, and R. Suslick, "Parameter-Independent Segmentation and Labeling of Speech," in Erman (ed.), 1974.

[Hal62] M. Halle and K. Stevens, "Speech Recognition: A Model and a Program for Research," *IRE Trans. PGIT*, IT-8, Pp. 155-159, 1962.

[Hay75] F. Hayes-Roth, and D. J. Mostow, "An Automatically Compilable Recognition Network for Structured Patterns," *Proc. Fourth Int. Joint Conf. on Artificial Intelligence*, Tbilisi, USSR, 1975.

[Hes74] W. J. Hess, "A Pitch Synchronous, Digital Feature Extraction System for Phonetic Recognition of Speech," in Erman (ed.), 1974.

[Ich73] A. Ichikawa, Y. Nakano, and K. Nakata, "Evaluation of Various Parameter Sets in Spoken Digits Recognition," *IEEE Trans. Audio and Electroacoustics*, AU-21 No.3, Pp.202-209, 1973.

[Ita68] F. Itakura and S. Saito, "Analysis Synthesis Telephony Based on the Maximum Likelihood Method," in *Proc. Sixth Int. Congr. Acoustics*, Paper C-5-5, 1968.

[Ita70] F. Itakura and S. Saito, "A Statistical Method for Estimation of Speech Spectral Density and Formant Frequencies," *Electron. Commun. Japan*, V-53-A, Pp.36-43, 1970.

[Ita75] F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition, *IEEE Trans. ASSP*, 23 67-72, 1975.

[Kam75] I. Kameny, "Comparison of the Formant Spaces of Retroflexed and Nonretroflexed Vowels," *IEEE Trans. ASSP, 23*, 38-49, 1975.

[Kla75] D. H. Klatt, "On the Design of Speech Understanding Systems," in G. Fant (ed.), *Proceedings of the Speech Communications Seminar*, Wiley, New York, 1975.

[Kla72] D. H. Klatt and K. N. Stevens, "Sentence Recognition from Visual Examination of Spectrograms and Machine Aided Lexical Searching," 1972 International Conference on Speech Communications and Processing, Boston, April, 1972.

[Kla73] D. H. Klatt and K. N. Stevens, "On the Automatic Recognition of Continuous Speech: Implications of a Spectrogram-Reading Experiment," *IEEE Trans. AU*, 21, 210-217, 1973.

[Kri75] J. S. Kriz, "A 16-Bit A-D-A Conversion System for High-Fidelity Audio Research," *IEEE Trans. ASSP*, 23, 146-148, 1975.

[Lad69] P. Ladefoged, *Three Areas of Experimental Phonetics*, Oxford University Press, London, 1969.

[Les75] V. R. Lesser, R. D. Fennell, L. D. Erman, and D. R. Reddy, "Organization of the Hearsay-II Speech Understanding System," *IEEE Trans. ASSP*, 23, 11-23, 1975.

[Lic64] J. C. R. Licklider, "Theory of Signal Detection," In Swets (ed.), 1964.

[Low76] B. Lowerre, A Comparative Performance Analysis of Speech Understanding Systems," Ph.D. Thesis (in preparation), Computer Science Dept., Carnegie-Mellon University, Pittsburgh, Pa., 1976.

[Lyo69] J. Lyons, *Introduction of Theoretical Linguistics*, Cambridge University Press, 1969.

[Mak75] J. Makhoul, "Linear Prediction: A Tutorial Review," *Proc. IEEE Special Issue on Digital Signal Processing*, 63, 4, 561-580, 1975.

[Mak72] J. I. Makhoul and J. J. Wolf, "Linear Predication and the Spectral Analysis of Speech," BBN report 2304, Cambridge, 1972.

[Man68] O. L. Mangasarian, "Multisurface Method of Pattern Separation," *IEEE Trans. Information Theory*, 14, 801-807, 1968.

[Mar72] J. D. Markel, "Digital Inverse Filtering: A New Tool for Formant Trajectory Estimation," *IEEE Trans. AU*, 20, 129-137, 1972.

[Mel72] W. S. Meisel, *Computer-oriented Approaches to Pattern Recognition*, Academic Press, N.Y., 1972.

[Mer75] P. Mermelstein, "A Phonetic-Context Controlled Strategy for Segmentation and Phonetic Labeling of Speech," *IEEE Trans. ASSP*, 23, 79-82, 1975.

[Min63] M. Minsky, "Steps Towards Artificial Intelligence," in Feigenbaum (ed.), Pp. 406-450, 1963.

[Mol74] L. Molho, "Automatic Recognition of Fricatives and Plosives in Continuous Speech," in Erman (ed.), 1974.

[Nag68] G. Nagy, "State of the Art in Pattern Recognition," *Proc. IEEE*, 56, No. 5, 836-862, May 1968.

[Nag66] G. Nagy and G. L. Shelton Jr., "Self-corrective Character Recognition System," *IEEE Trans. Information Theory*, IT-12, 215-222, Apr. 1966.

[Nel68] G. D. Nelson, and D. M. Levy, "A Dynamic Programming Approach to the Selection of Pattern Features," *IEEE Trans. System Science and Cybernetics*, 4, 145-151, 1968.

[New71] A. Newell, J. Barnett, J. Forgie, C. Green, D. Klatt, J. C. R. Licklider, J. Munson, R. Reddy, and W. Woods, *Speech Understanding Systems: Final Report of a Study Group*, 1971. Reprinted by North-Holland/American Elsevier, Amsterdam, 1973.

[New75] A. Newell, "A Tutorial on Speech Understanding Systems," in Reddy (ed.), 1975.

[Ohl75] R. Ohlander, "Analysis of Natural Scenes," PhD. Thesis, Computer Science Dept., C-MU, Pittsburgh, 1975.

[Old68] R. C. Oldfield and J. C. Marshall (eds.), *Language: Selected Readings*, Penguin, Baltimore, 1968.

[OppG8] A. V. Oppenheim and R. W. Schafer, "Homomorphic Analysis of Speech," *IEEE Trans. Audio and Electroacoustics*, AU-16, No. 2, Pp. 221-226, 1968.

[Pat69] E. A. Patrick, and F. P. Fischer II, "Nonparametric Feature Selection," *IEEE Trans. Information Theory*, 15, 577-584, 1969.

[Red66] D. R. Reddy, "Segmentation of Speech Sounds," *J. Acoustic Society Amer.*, 40, 307-312, 1966.

[Red67] D. R. Reddy, "Phoneme Grouping for Speech Recognition," *J. Acoust. Soc. Amer.*, 41, No. 5, 1295-1300, May 1967.

[Red75a] D. R. Reddy (ed.), *Speech Recognition: Invited Papers of the IEEE Symposium*, Academic Press, NY, 1975.

[Red75b] D. R. Reddy, "Speech Recognition by Machine: A Review," to be published, *Proc. IEEE*, April 1976.

[Red73] D. R. Reddy, L. D. Erman, and R. B. Neely, "A Model and a System for Machine Recognition of Speech," *IEEE Trans. AU*, 21, 229-238, 1973.

[Rit74] H. B. Ritea, "A Voice-Controlled Data Management System," in Erman (ed.), 1974.

[Ros57] F. Rosenblatt, "The Perceptron -- A Perceiving and Recognizing Automaton," Rept. 85-460-1, Cornell Aerodynamical Lab., Ithaca, N.Y., Jan. 1957.

[SchRW75] R. W. Schafer and L. R. Rabiner, "Parametric Representations of Speech," in Reddy (ed.), 1975.

[SchR75] R. Schwartz, and J. Makhoul, "Where the Phonemes Are: Dealing with Ambiguity in Acoustic-Phonetic Recognition," *IEEE Trans. ASSP*, 23, 50-53, 1975.

[Sel63] O. G. Selfridge and U. Neisser, "Pattern Recognition by Machine," in Feigenbaum (ed.), Pp. 237-250, 1963.

[Sho74a] L. Shockey and R. Reddy, "Quantitative Analysis of Speech Perception: Results from Transcription of Continuous Speech from Unfamiliar Languages," Computer Science Dept., C-MU, Pittsburgh, 1974. (paper presented at: Speech Communications Seminar, Stockholm, August, 1974.)

[Sho74b] L. Shockey, "Description of CMU Allophone Sentences," SUR note N22188, 1974.

[Sil75] H. F. Silverman and N. R. Dixon, "An Objective Parallel Evaluator of Segmentation/Classification Performance for Multiple Systems," *IEEE Trans. ASSP*, 23, 92-99, 1975.

[SmiAR75] A. R. Smith, "POMOW: Hearsay-II Word Hypothesizer," Tech. Report, Computer Science Department, Carnegie-Mellon University, Pittsburgh, Pa., 1975.

[SmiFM68] F. W. Smith, "Pattern Classification Design by Linear Programming," *IEEE Trans. Computers*, 17, 367-372, 1968.

[Tap71] C. C. Tappert, M. R. Dixon, A. S. Rabinowitz and W. D. Chapman, "Automatic Recognition of Continuous Speech Utilizing Dynamic Segmentation, Dual Classification, Sequential Decoding and Error Recovery," Rome Air Development Center, Griffiss AFB, Rome, N.Y., Tech Rep., TR-71-146, 1971.

[Tan64] W. P. Tanner and T. B. Birdsall, "Definitions of d' and (eta) as Psychophysical Measures," in Swets (ed.), 1964.

[Uhr63] L. Uhr and C. Vossler, "A Pattern Recognition Program that Generates, Evaluates, and Adjusts its own Operators," in Feigenbaum (ed.), Pp. 251-268, 1963.

[Uhr66] L. Uhr, "Pattern Recognition," in: *Pattern Recognition*, 365-381, John Wiley & Sons, N.Y., 1966.

[Uhr73] L. Uhr, *Pattern Recognition, Learning, and Thought*, Prentice Hall Inc., Englewood Cliff, N.J., 1973.

[Vic69] P. Vicens, "Aspects of Speech Recognition by Computer," Rept. CS-127, Ph.D. Thesis, Computer Science Department, Stanford Univ., 1969.

[Wal47] A. Wald, *Sequential Analysis*, John Wiley & Sons, N.Y., 1947.

[Wei75] C. J. Weinstein, S. S. McCandless, L. F. Mondshein and V. W. Zue, "A System for Acoustic-Phonetic Analysis of Continuous Speech," *IEEE Trans. ASSP*, 23, 54-67, 1975.

[Whi75] G. M. White and R. B. Neely, "Speech Recognition Experiments with Linear Prediction, Bandpass Filtering and Dynamic Programming," *Proc. Second USA-Japan Computer Conference*, Tokyo, Japan, August 1975, also to appear in *IEEE Trans. ASSP*, 23, December 1975.

[Wie74] R. A. Wiesen and J. W. Forgie, "An Evaluation of the Lincoln Laboratory Speech Recognition System," *J. Acoustic Society Amer.*, 56, S27 (A), 1974.

[Wil68] M. B. Wilk and R. Gnanadesikan, "Probability Plotting Methods for the Analysis of Data," *Biometrika* Vol. 55, No. 1, P. 1, 1968.

[Woo75] W. A. Woods, "Motivation and Overview of SPEECHLIS: An Experimental Prototype for Speech Understanding Research," *IEEE Trans. ASSP*, 23, 2-10, 1975.

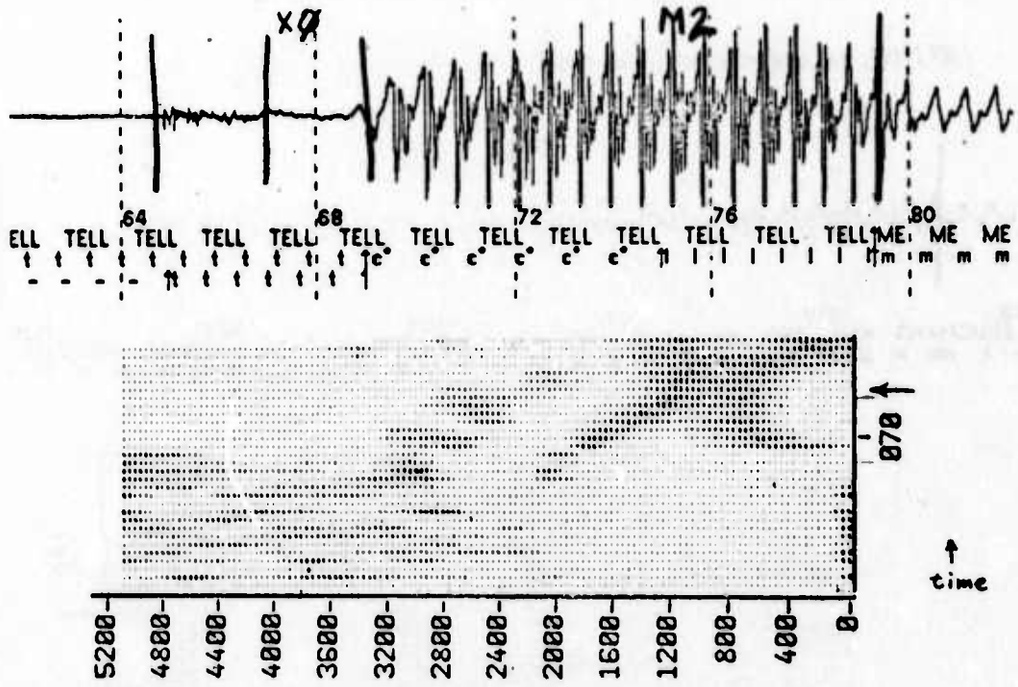
[Woo74] W. A. Woods, M. A. Bates, B. C. Bruce, J. J. Colarusso, C. C. Cook, L. Gould, J. I. Makhoul, B. L. Nash-Webber, R. M. Schwartz, and J. J. Wolf, "Speech Understanding Research at BBN, Final Report on Natural Communication with Computers," Volume I, Bolt Beranek and Newman Inc., Report 2976, 1974.

S1: Segmentation -- Some Cases

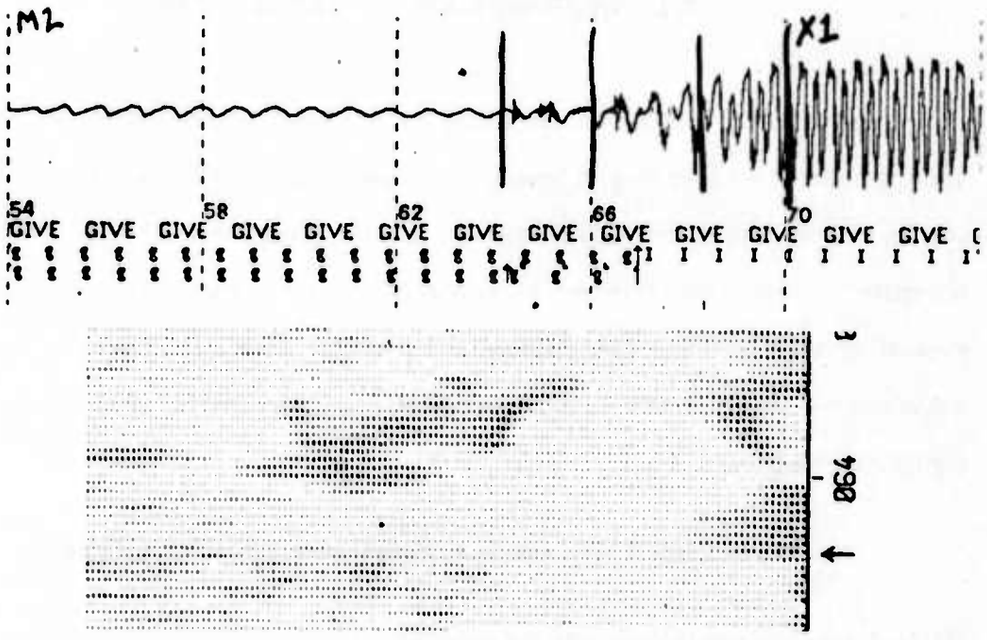
The following are some cases where the hand and machine segmentations disagree. They are classified according to type of error ((M)issing or e(X)tra) and degree (0-machine correct, 1-not critical, 2-critical error). We introduce these cases to illustrate the various phenomena which are involved in segmentation, and which must be considered in evaluating segmentation. Two displays are given for each case: a plot of the digitized waveform -- 10kHz., 9 bits -- and a plot of the SPG parameters (which serves well as a digital spectrogram).

M2 -- Cases where a critical segment boundary is not detected:

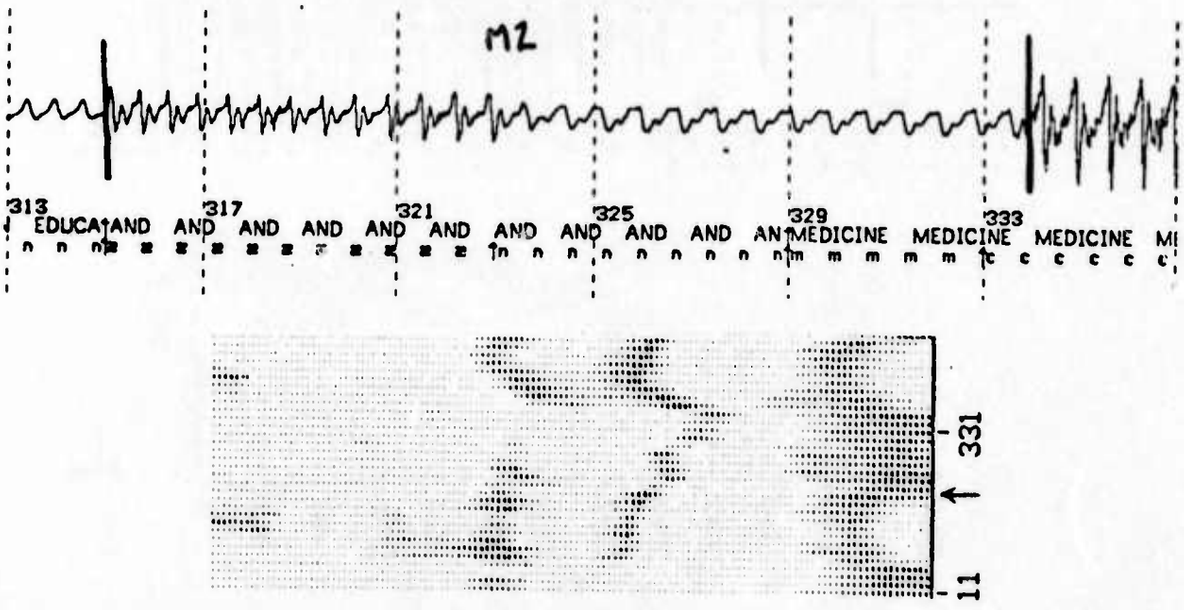
/EH//L/, slow change in sonorants not detected



/-//G/, voice bar not detected

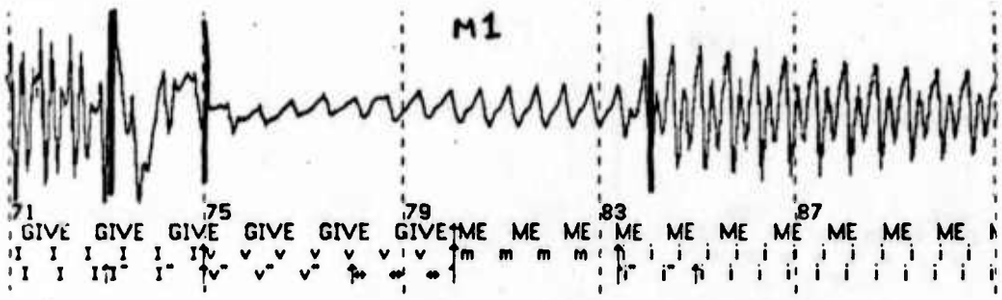


/AE//N/, nasalized vowel and nasal

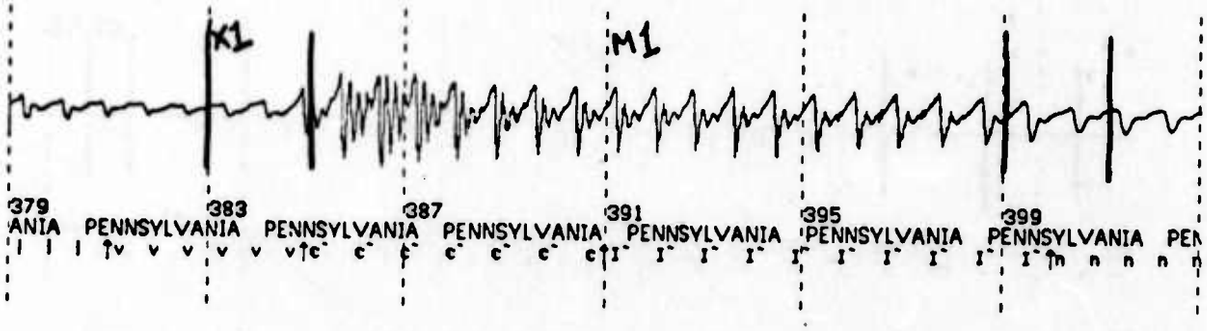


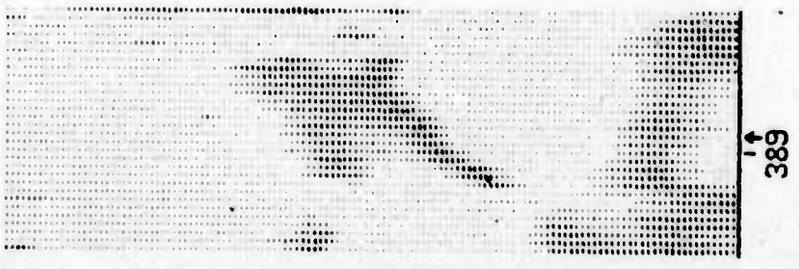
M1 -- Cases where a non-critical boundary is missed:

/V//M/, very slight change, phonologically explainable



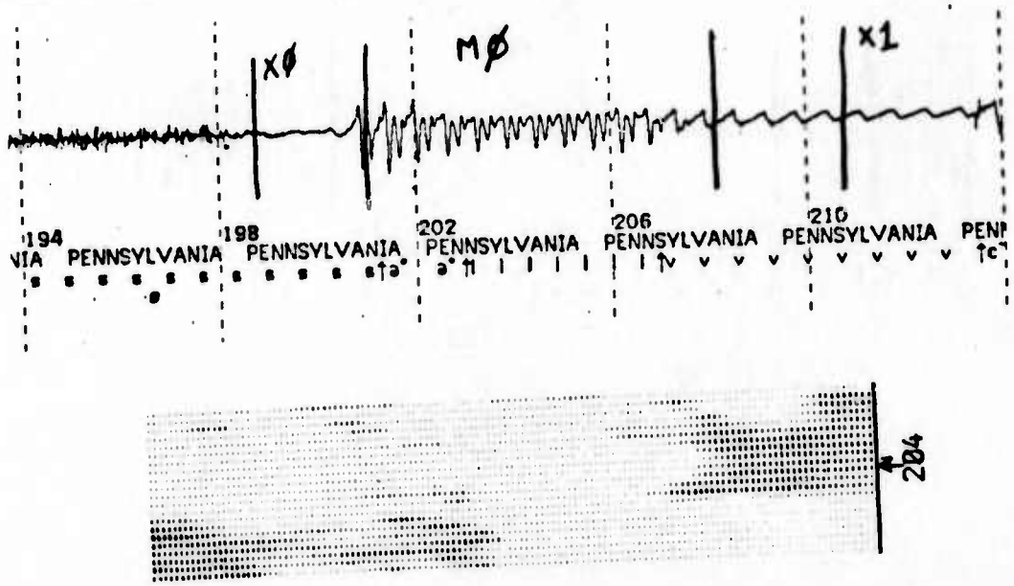
/EH//IH/, diphthong, similar sounds, often merged



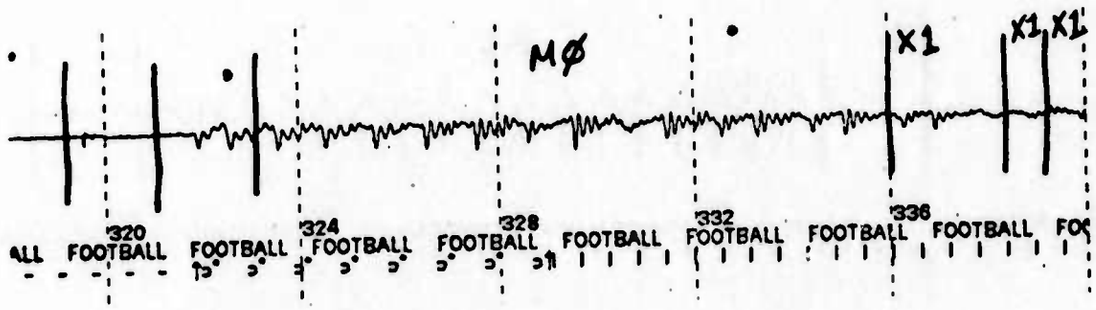


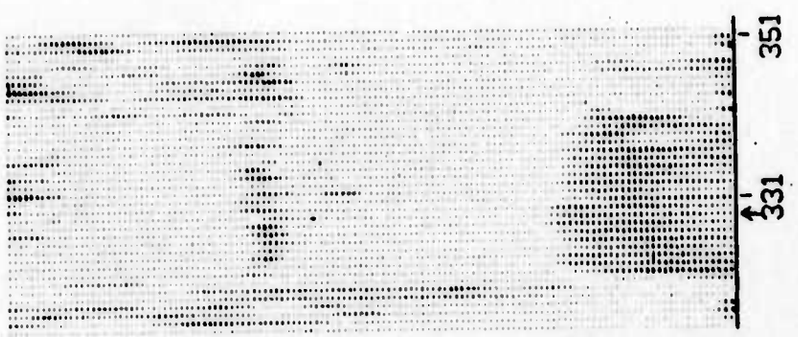
MO -- Cases where hand segmentation is not correct

/EL/, no separate vowel segment

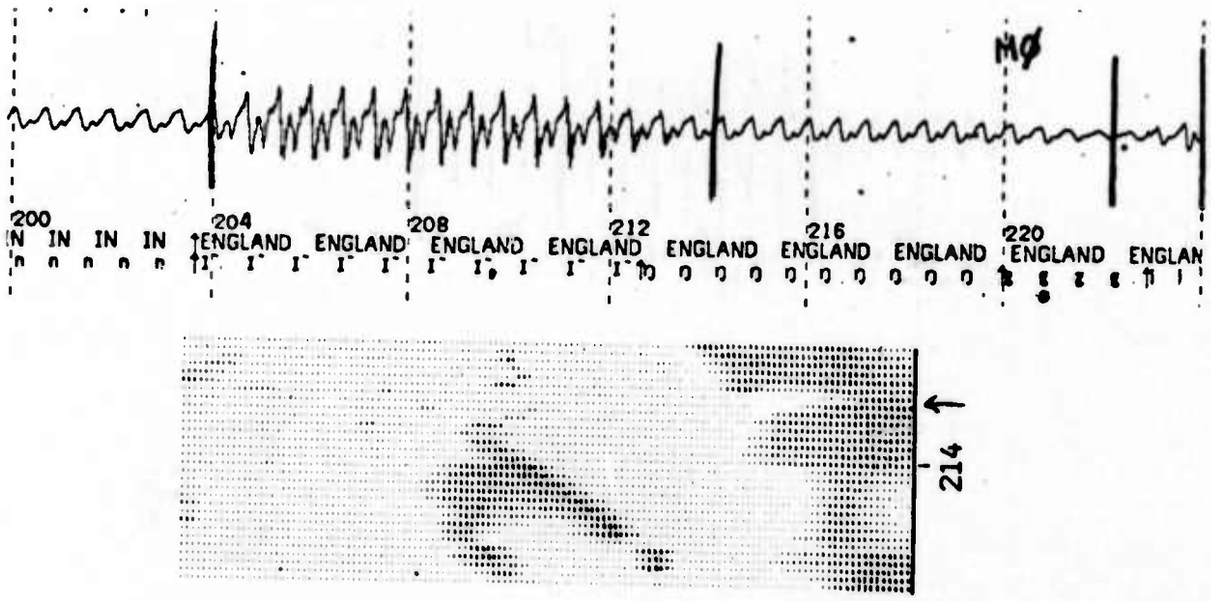


/AO/, utterance-final, no separate /L/



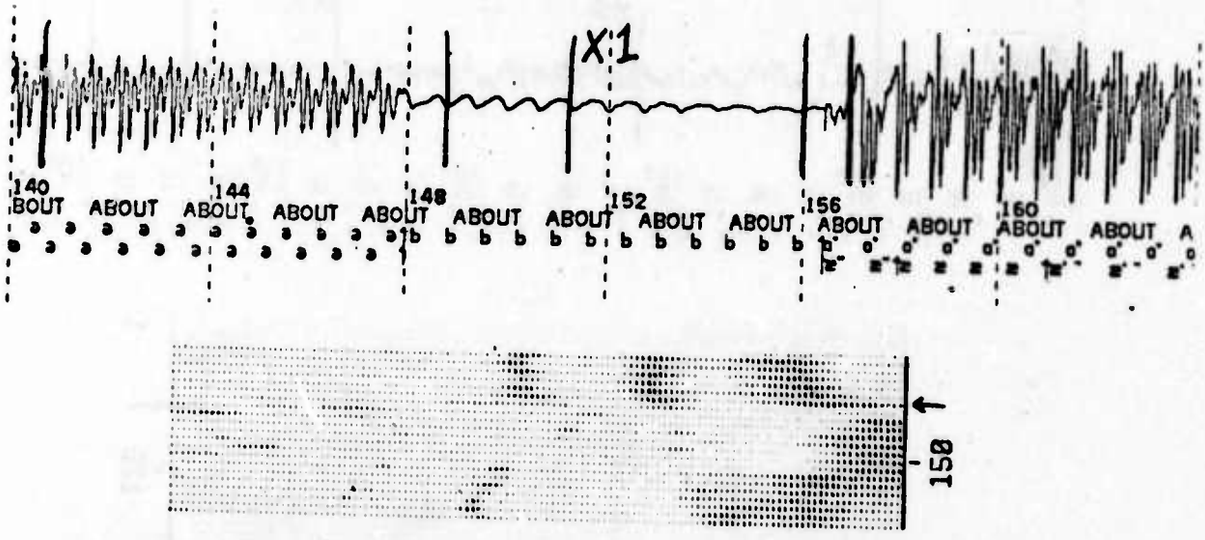


/NG/, nasal to voice bar, /G/ probably deleted

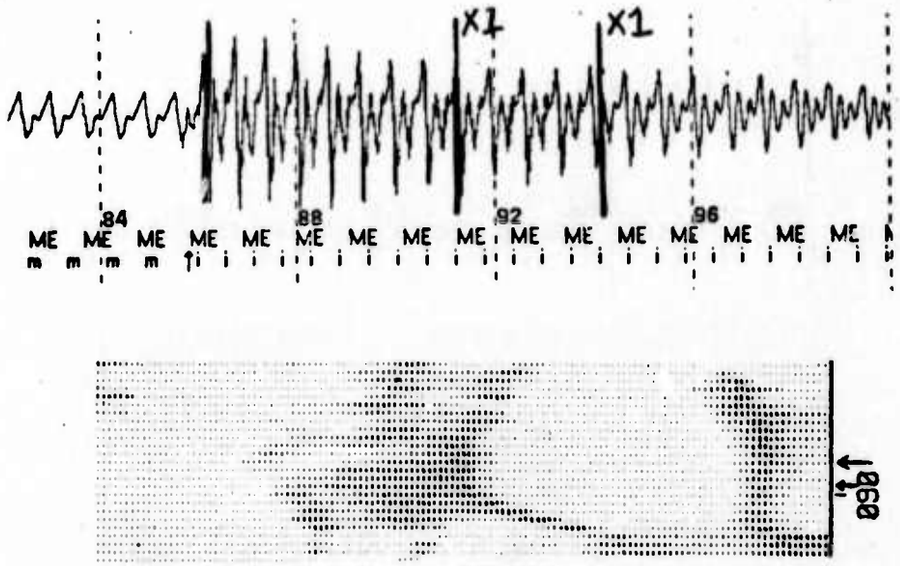


X1 -- Cases where machine boundary is incorrectly included:

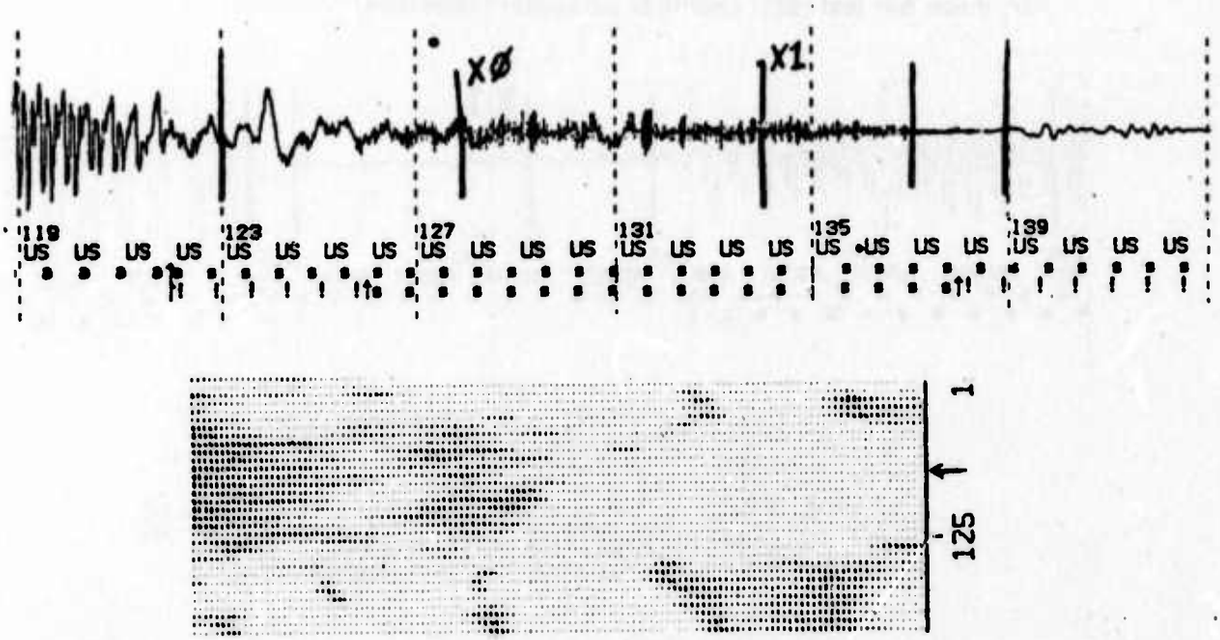
/B/, voice bar lost (SPG amplitude parameter insensitive)



/IY/, vowel segment broken up

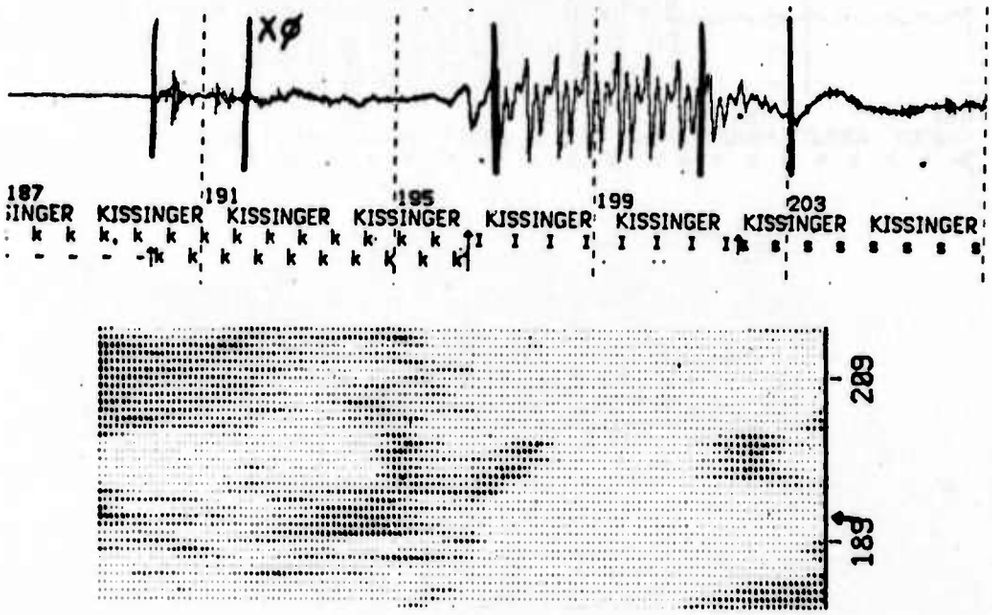


/S/, fricative broken up (133), note (128) correct

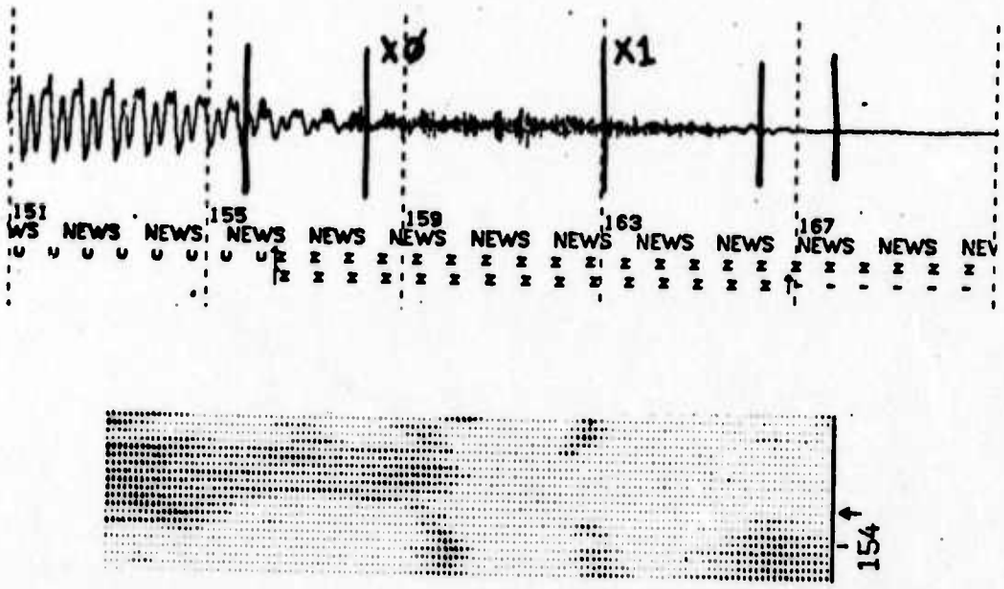


X0 -- Cases where hand boundary should be indicated:

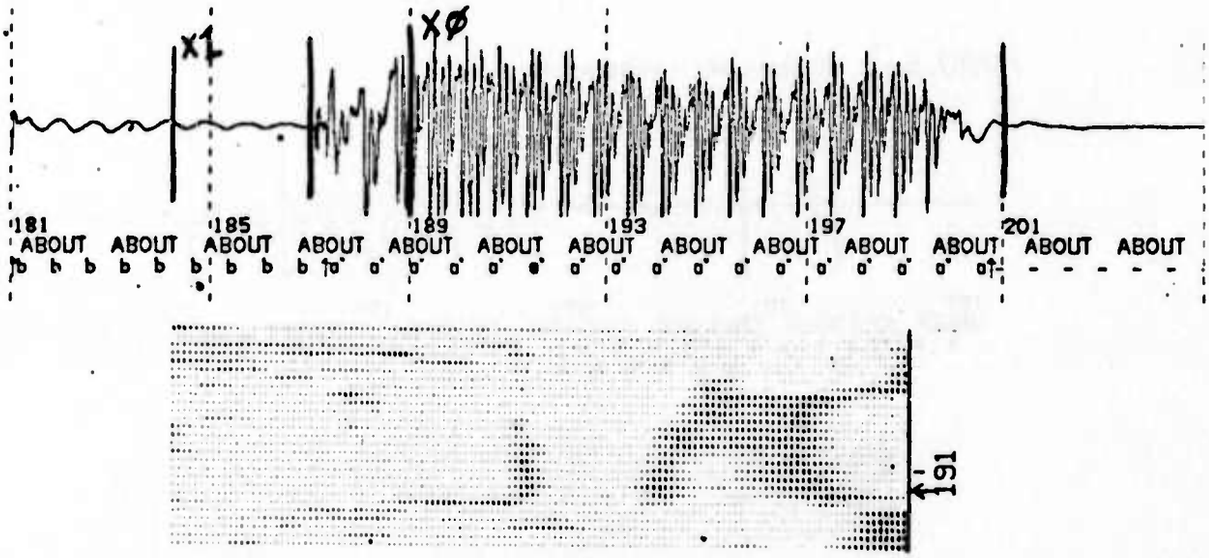
/K//K/, burst and aspiration separated



/Z//S/, voiced transition segment detected

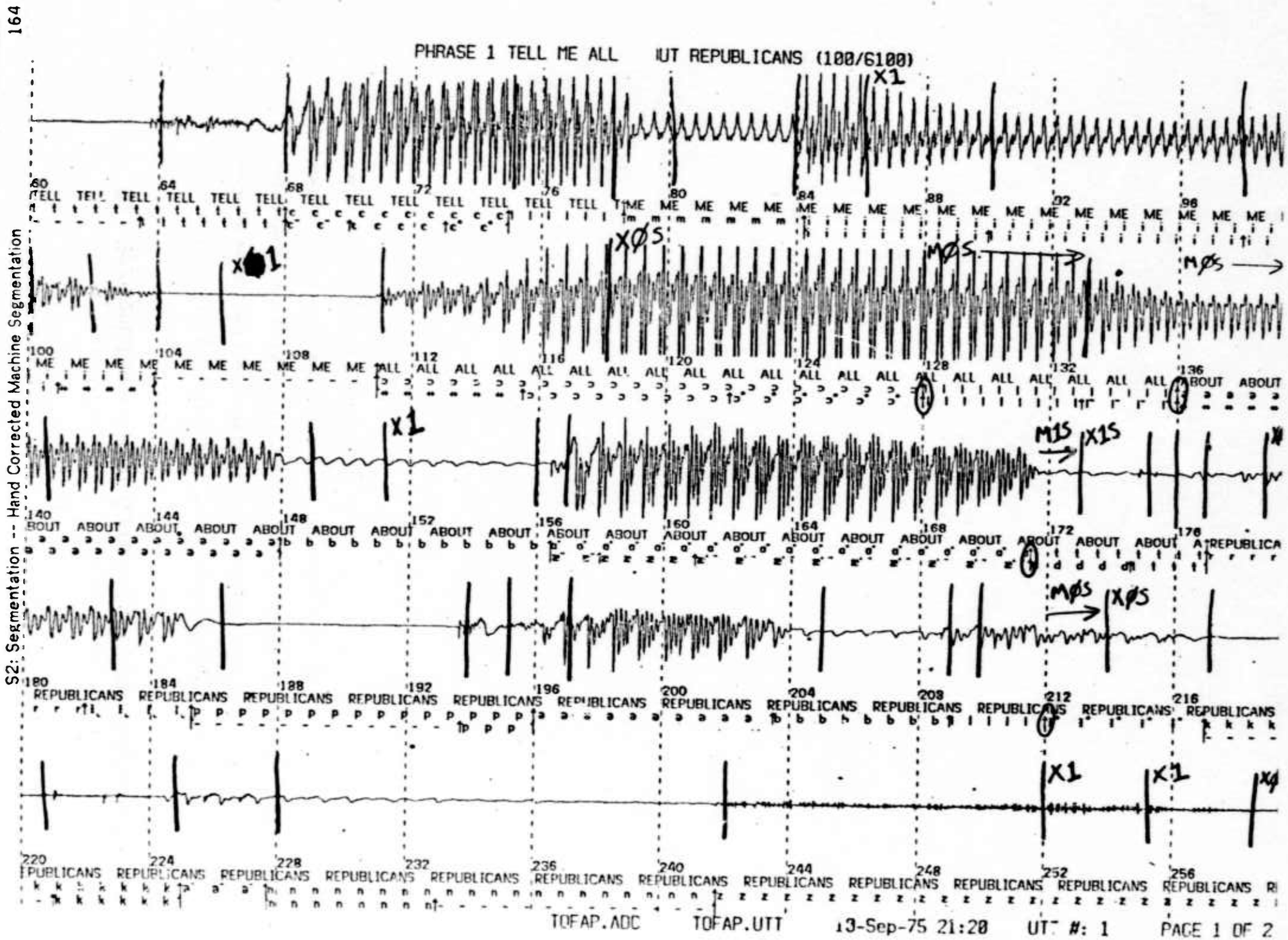


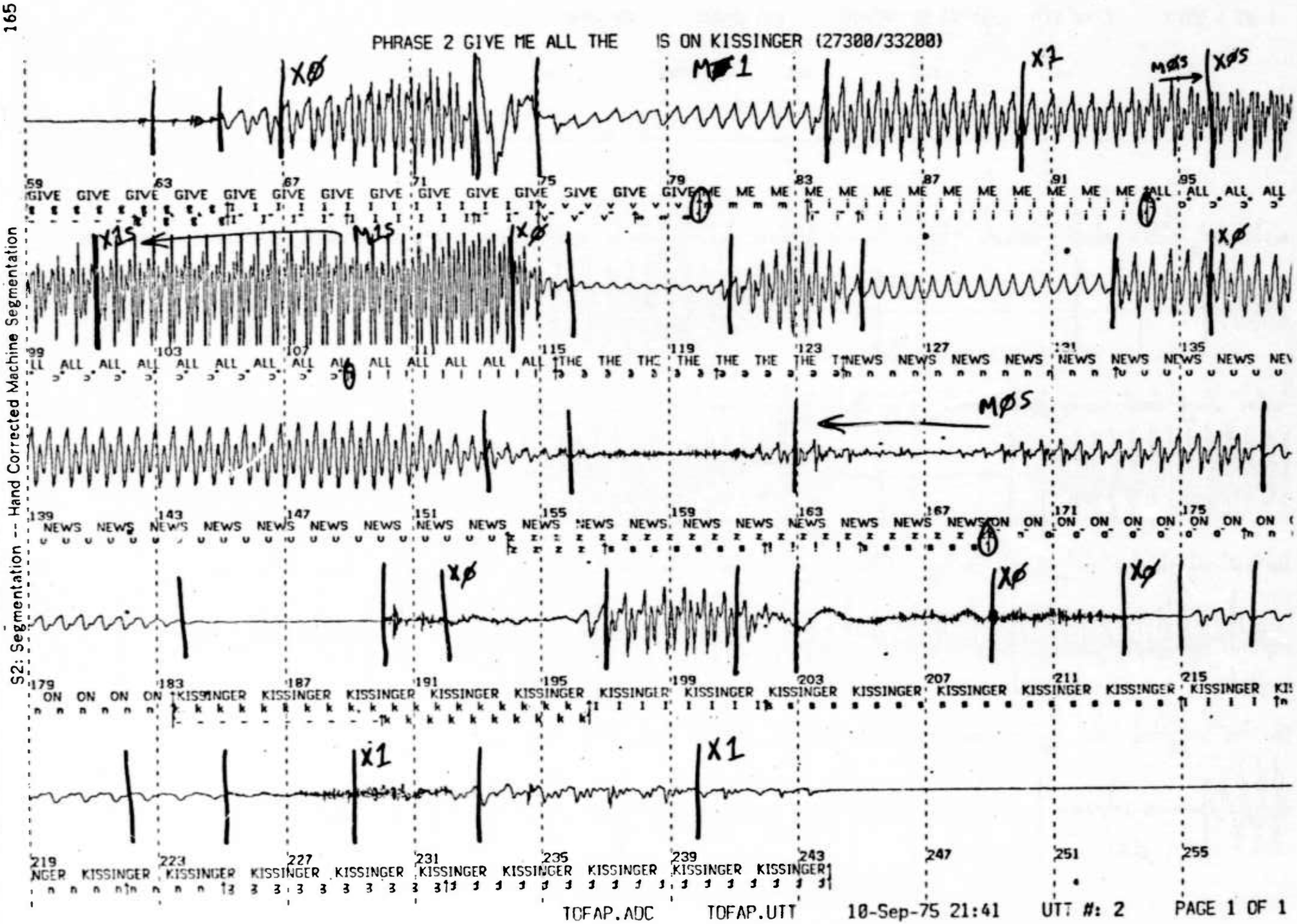
/AW//AW/, vowel onset transition segment

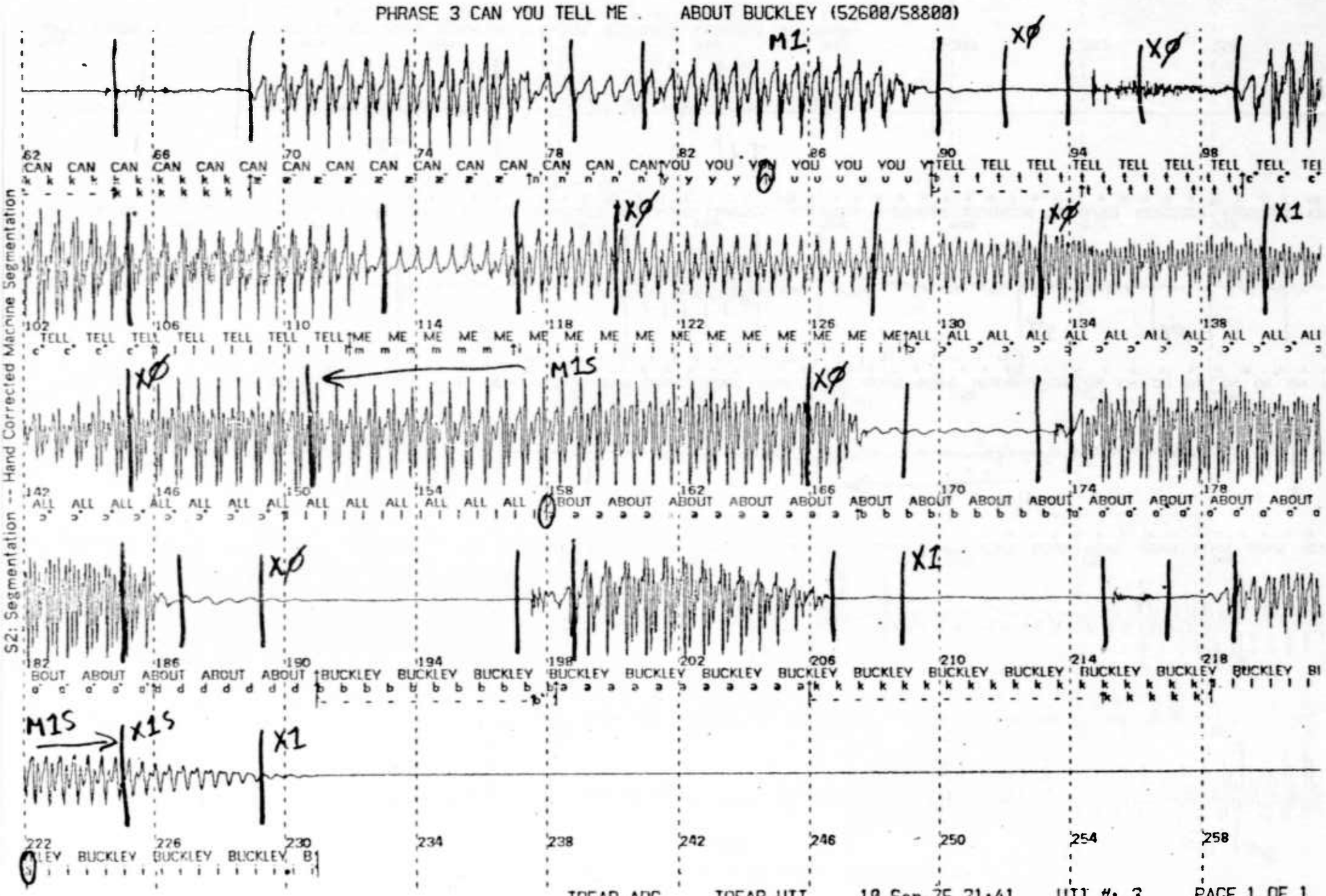


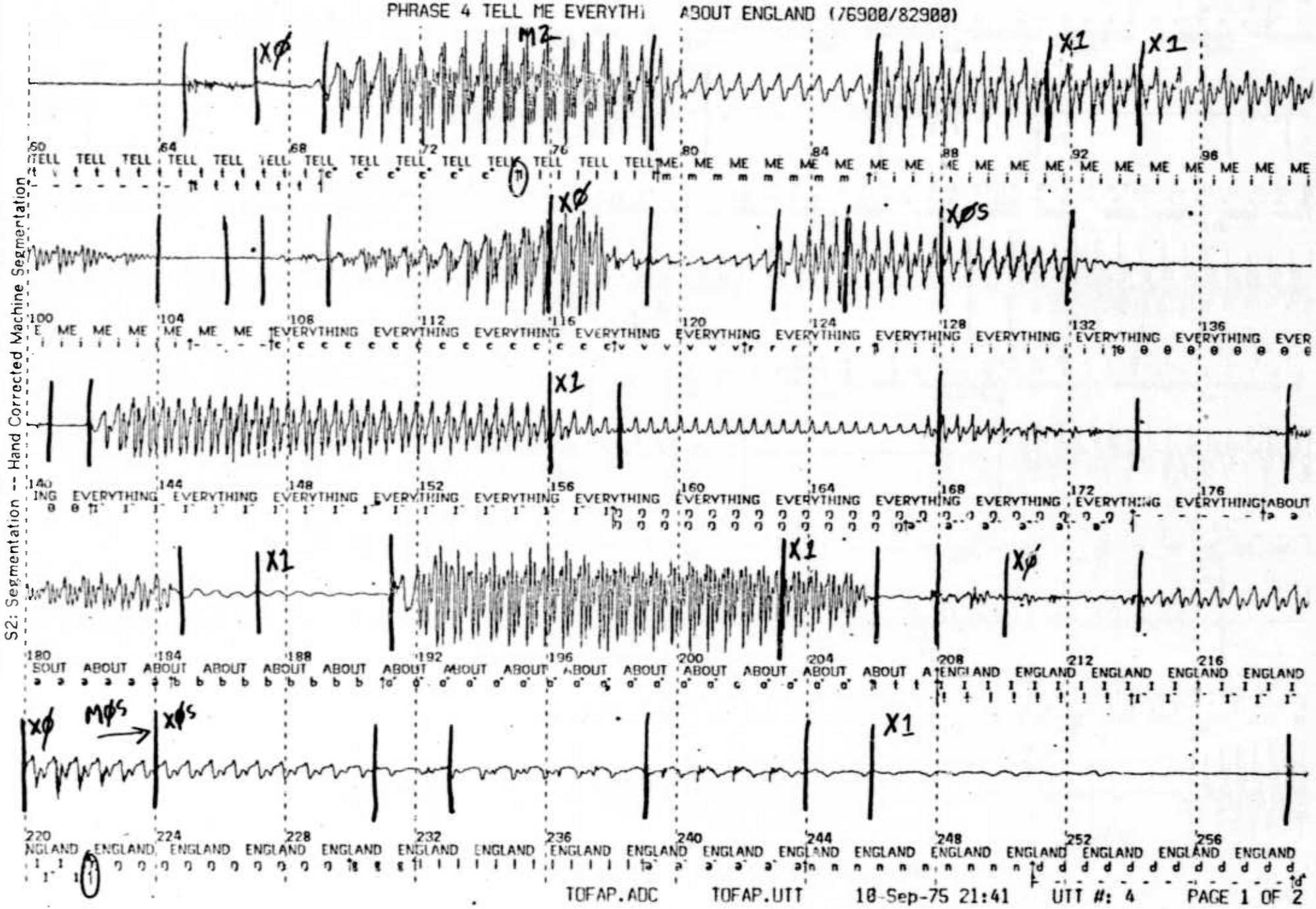
S2: Segmentation -- Hand Corrected Machine Segmentation

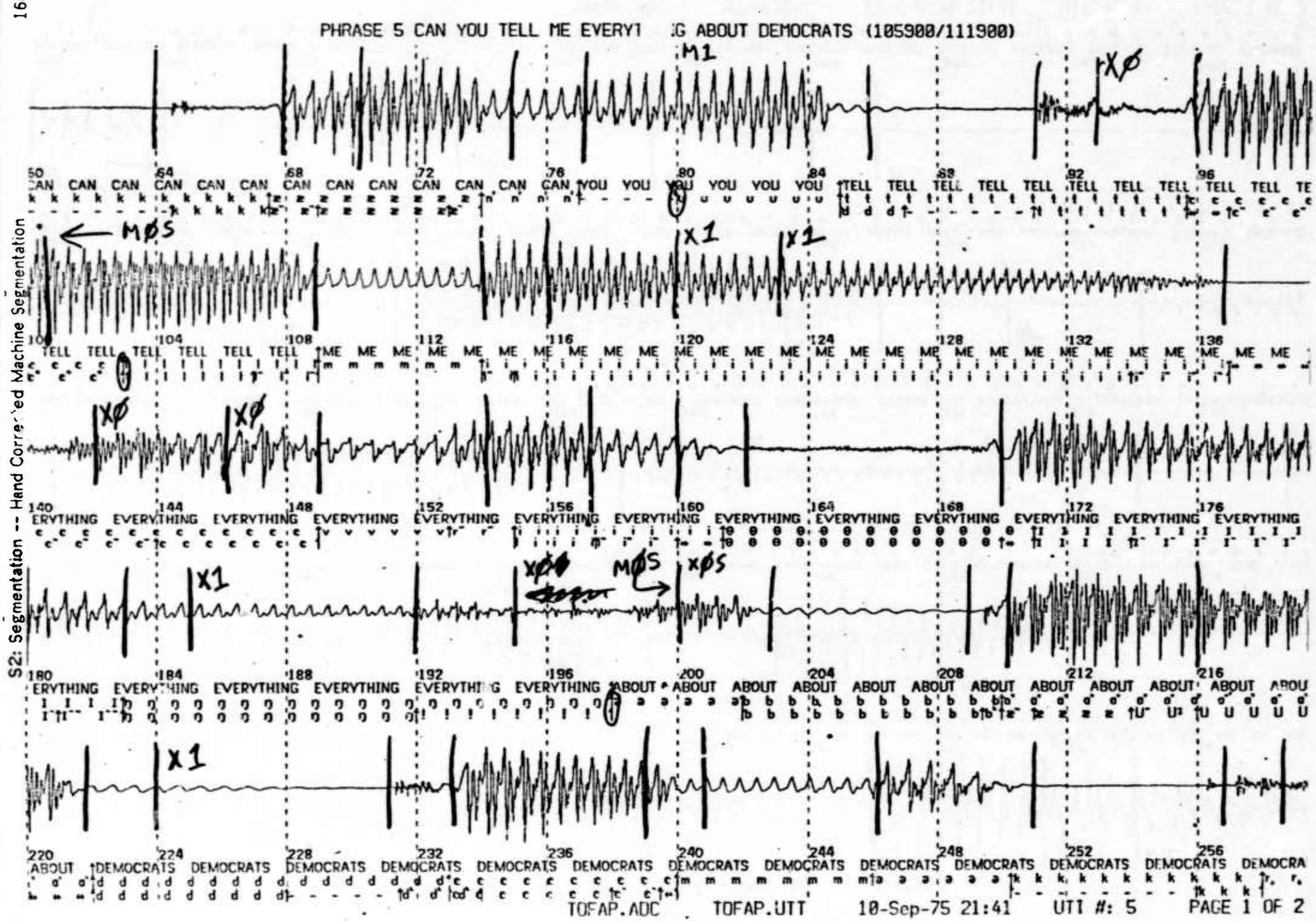
In the following waveform plots, the results of running the segmenter with SPG input parameters are shown. Ratings of all the points of disagreement with the referent segmentation are given. In this plot, the referent segmentation is given below the waveform in two lines, a phonemic and a sub-phonemic transcription. The ratings are the same types as indicated in Appendix S1.

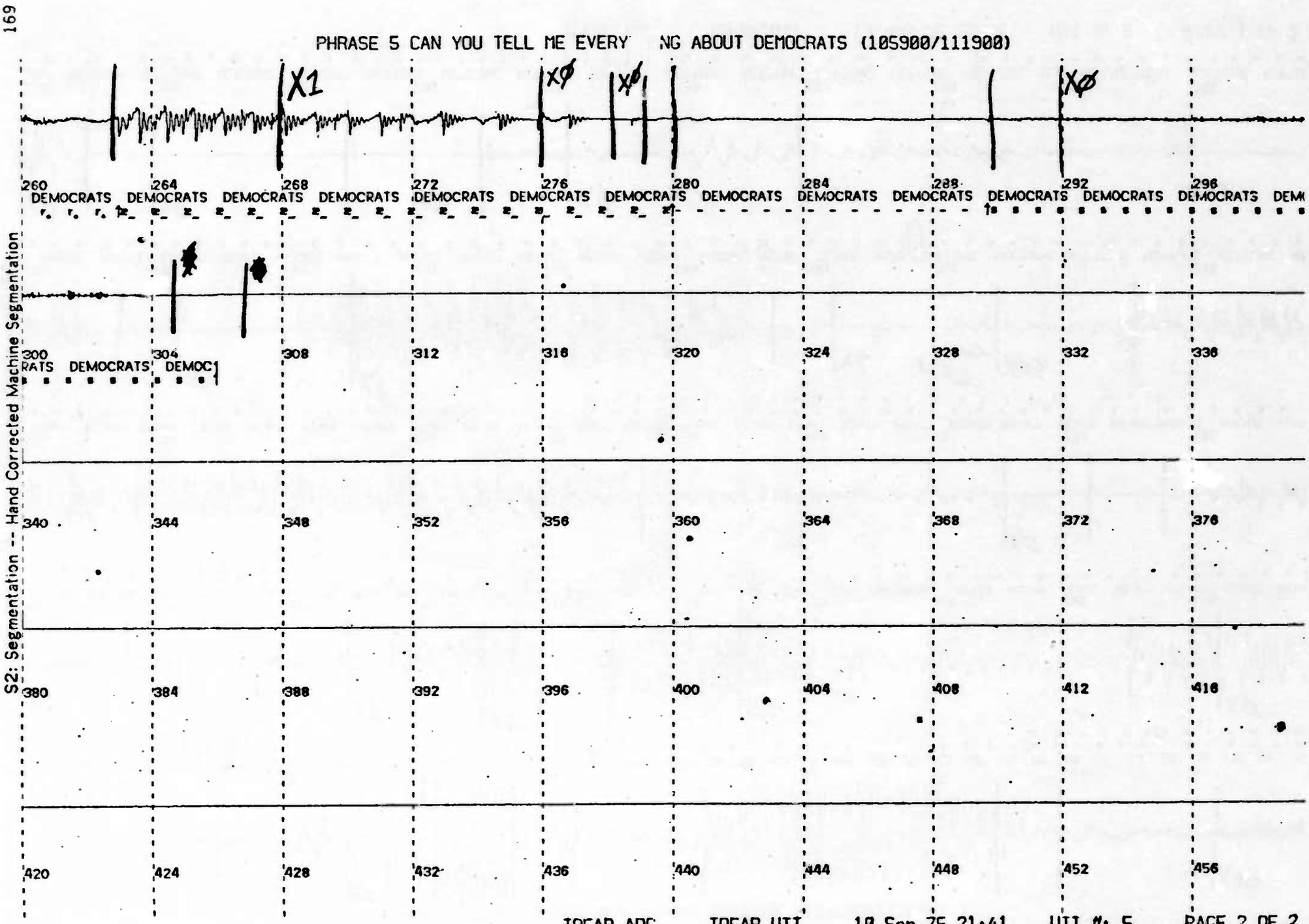

PHRASE 1 TELL ME ALL UT REPUBLICANS (100/6100)
S2: Segmentation -- Hand Corrected Machine Segmentation
TOFAP.ADC TOFAP.UTT 13-Sep-75 21:20 UTT #: 1 PAGE 1 OF 2


PHRASE 2 GIVE ME ALL THE IS ON KISSINGER (27300/33200)
S2: Segmentation -- Hand Corrected Machine Segmentation
X0
M1
X1
M0S
X0S
M1S
59 GIVE GIVE GIVE GIVE GIVE GIVE GIVE GIVE GIVE GIVE GIVE GIVE GIVE ME ME ME ME ME ME ME ME ME ME ALL ALL ALL ALL
99 ALL ALL ALL ALL ALL ALL ALL ALL ALL ALL ALL ALL ALL THE THE THE THE THE THE THE NEWS NEWS NEWS NEWS NEWS NEWS NEWS NEWS NEW
139 NEWS NEWS NEWS NEWS NEWS NEWS NEWS NEWS NEWS NEWS NEWS NEWS NEWS NEWS NEWS NEWS NEWS NEWS ON ON ON ON ON ON ON ON ON
179 ON ON ON ON KISSINGER KISSINGER KISSINGER KISSINGER KISSINGER KISSINGER KISSINGER KISSINGER KISSINGER KISSINGER KISSINGER KISSINGER KISSINGER KISSINGER KI
219 NGER KISSINGER KISSINGER KISSINGER KISSINGER KISSINGER KISSINGER KISSINGER KISSINGER KISSINGER KISSINGER
TOFAP.ADC TOFAP.UTT 10-Sep-75 21:41 UTT #: 2 PAGE 1 OF 1

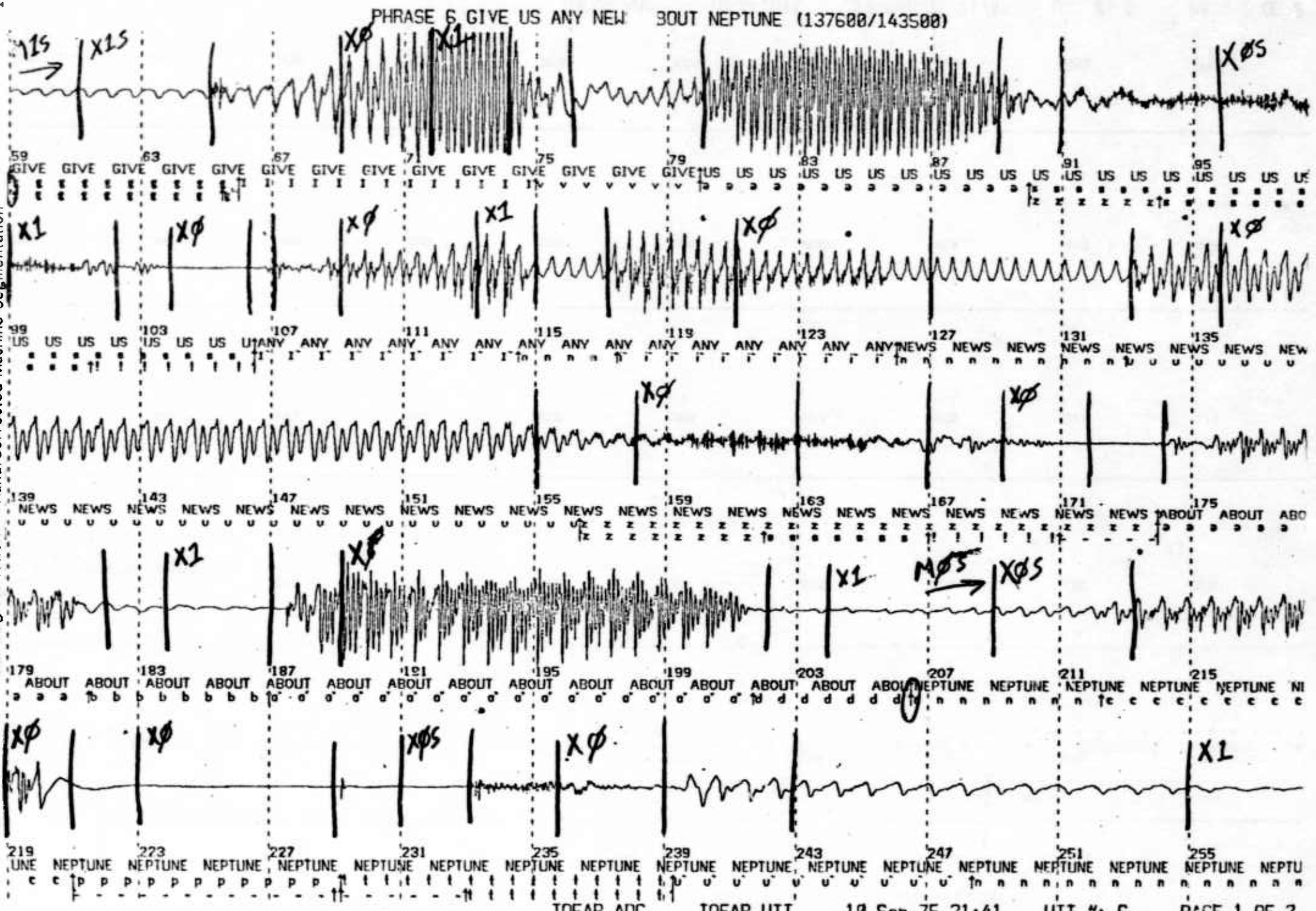

PHRASE 3 CAN YOU TELL ME . ABOUT BUCKLEY (52600/58800)
S2: Segmentation -- Hand Corrected Machine Segmentation
M1
XØ
X1
M1S
X1S
TOFAP.ADC
TOFAP.UTT
10-Sep-75 21:41
UTT #: 3
PAGE 1 OF 1

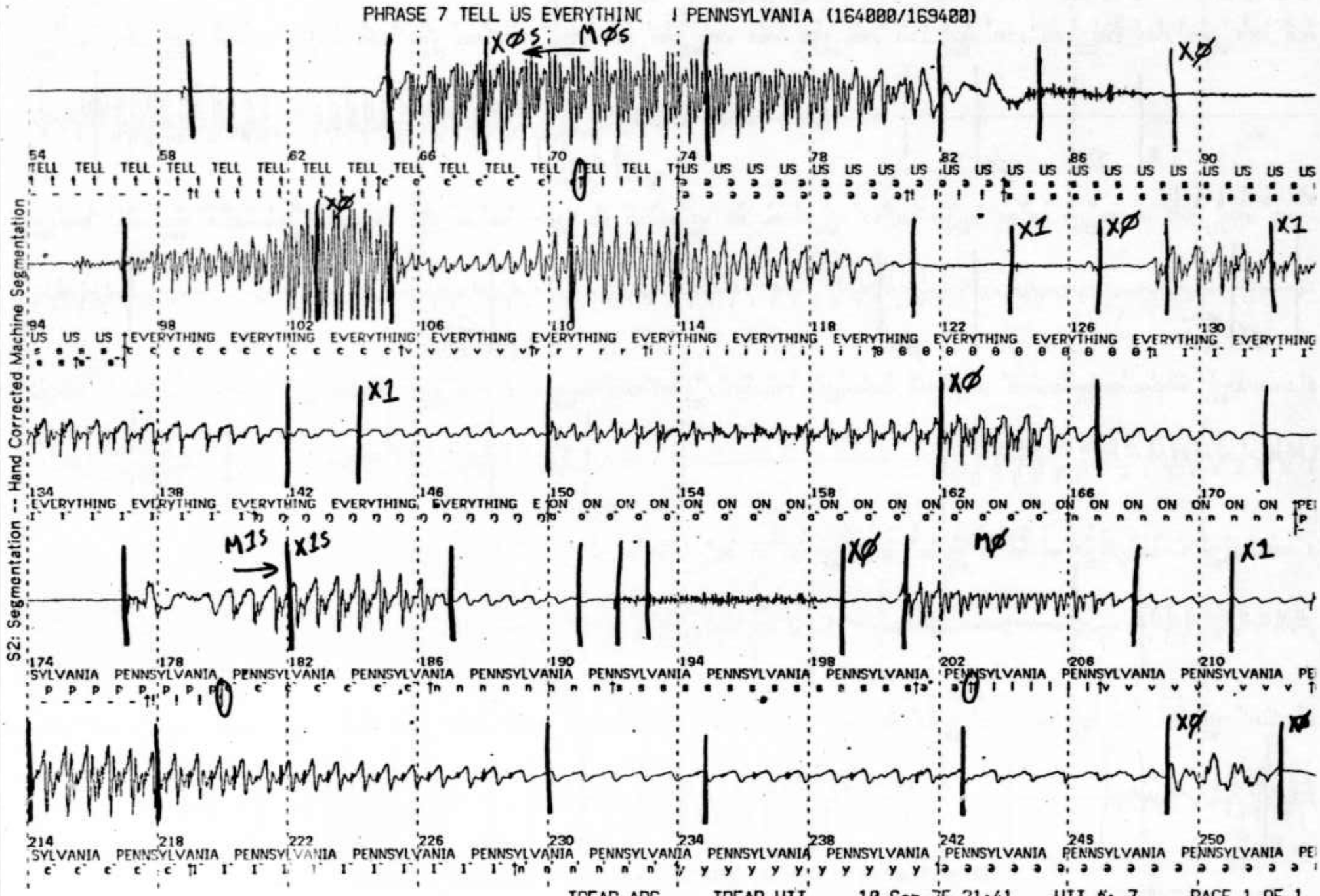

PHRASE 4 TELL ME EVERYTH ABOUT ENGLAND (76900/82900)
S2: Segmentation -- Hand Corrected Machine Segmentation
TOFAP.ADC TOFAP.UTT 18-Sep-75 21:41 UTT #: 4 PAGE 1 OF 2


PHRASE 5 CAN YOU TELL ME EVERYT G ABOUT DEMOCRATS (105900/111900)
S2: Segmentation -- Hand Corrected Machine Segmentation
M1
XØ
MØS
X1
X1
XØ
XØ
X1
XØ
MØS
XØS
X1
50 CAN CAN CAN CAN CAN CAN CAN CAN CAN CAN CAN CAN CAN YOU YOU YOU YOU YOU YOU TELL TELL TELL TELL TELL TELL TELL TELL TELL TE
100 TELL TELL TELL TELL TELL TELL ME ME ME ME ME ME ME ME ME ME ME ME ME ME ME ME ME ME ME ME ME ME ME ME ME
140 ERYTHING EVERYTHING EVERYTHING EVERYTHING EVERYTHING EVERYTHING EVERYTHING EVERYTHING EVERYTHING EVERYTHING EVERYTHING EVERYTHING EVERYTHING EVERYTHING
180 ERYTHING EVERYTHING EVERYTHING EVERYTHING EVERYTHING EVERYTHING ABOUT ABOUT ABOUT ABOUT ABOUT ABOUT ABOUT ABOUT ABOUT ABOUT ABOUT ABOU
220 ABOUT DEMOCRATS DEMOCRATS DEMOCRATS DEMOCRATS DEMOCRATS DEMOCRATS DEMOCRATS DEMOCRATS DEMOCRATS DEMOCRATS DEMOCRATS DEMOCRATS DEMOCRATS DEMOCRA
TOFAP.ADC TOFAP.UTT 10-Sep-75 21:41 UTT #: 5 PAGE 1 OF 2

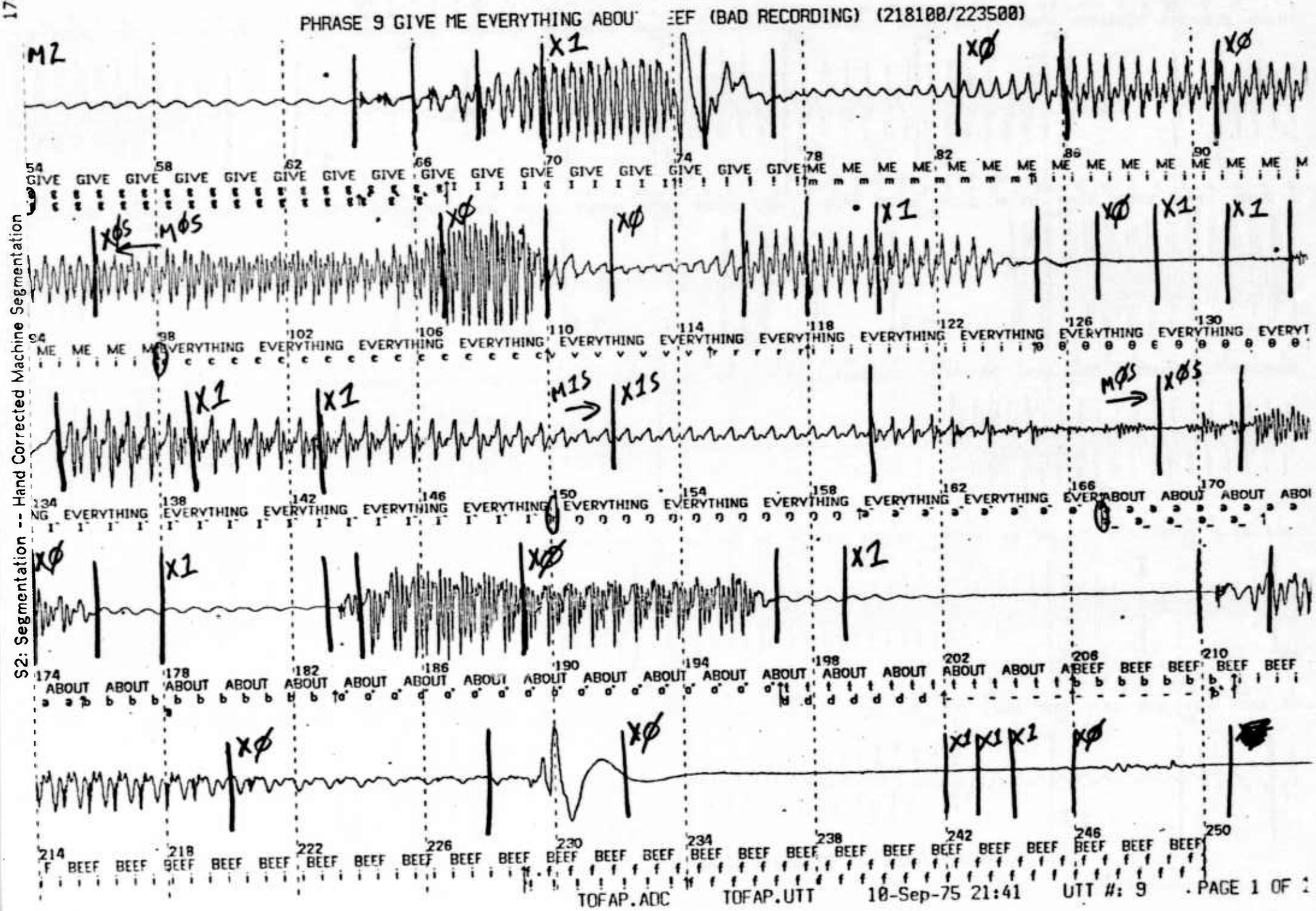

PHRASE 5 CAN YOU TELL ME EVERY NG ABOUT DEMOCRATS (105900/111900)
X1
XØ
XØ
XØ
260 264 268 272 276 280 284 288 292 296
DEMOCRATS DEMOCRATS DEMOCRATS DEMOCRATS DEMOCRATS DEMOCRATS DEMOCRATS DEMOCRATS DEMOCRATS DEMOCRATS DEMOCRATS DEMOCRATS DEMOCRATS DEM
300 304 308 312 316 320 324 328 332 336
RATS DEMOCRATS DEMOC
340 344 348 352 356 360 364 368 372 376
380 384 388 392 396 400 404 408 412 416
420 424 428 432 436 440 444 448 452 456
S2: Segmentation -- Hand Corrected Machine Segmentation

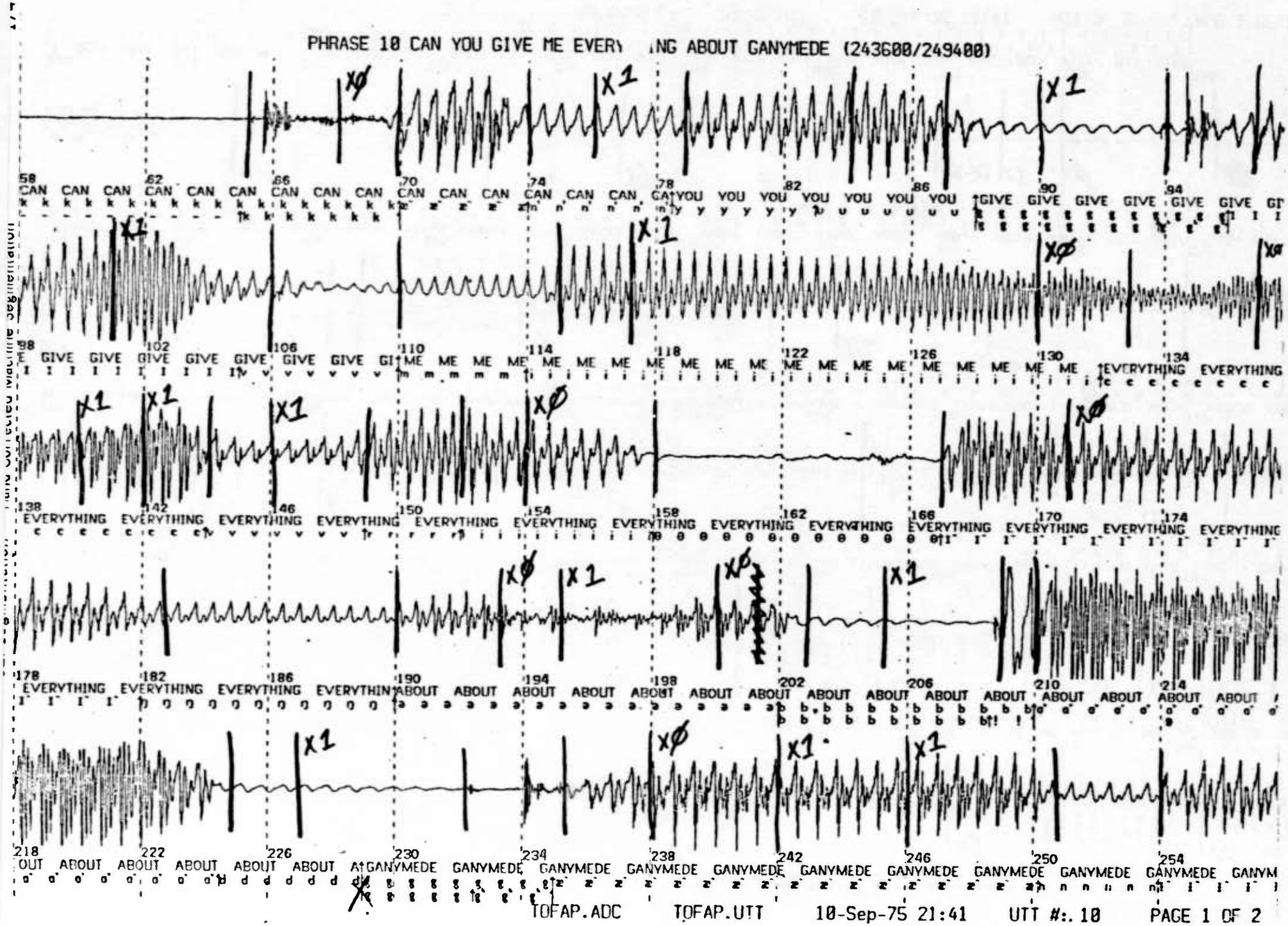
S2: Segmentation -- Hand Corrected Machine Segmentation




PHRASE 7 TELL US EVERYTHING I PENNSYLVANIA (164000/169480)
S2: Segmentation -- Hand Corrected Machine Segmentation
TOFAP.ADC. TOFAP.UTT 10-Sep-75 21:41 UTT #: 7 PAGE 1 OF 1

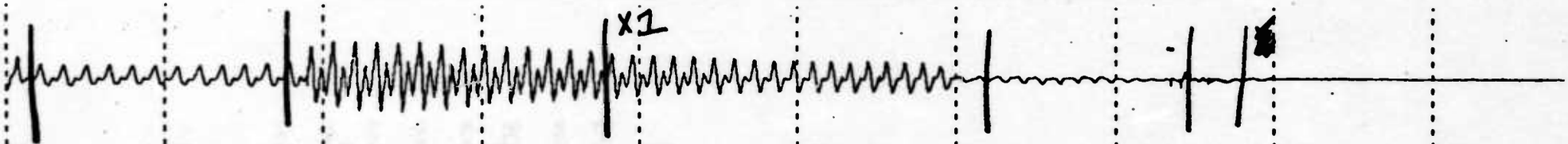

PHRASE 8 COULD YOU TELL US E YTHING ON GOL (190600/196600)
S2: Segmentation -- Hand Corrected Machine Segmentation
TOFAP.ADC TOFAP.UTT 10-Sep-75 21:41 UTT #: 8 PAGE 1 OF 2


PHRASE 9 GIVE ME EVERYTHING ABOU EF (BAD RECORDING) (218100/223500)
S2: Segmentation -- Hand Corrected Machine Segmentation
TOFAP.ADC TOFAP.UTT 10-Sep-75 21:41 UTT #: 9 .PAGE 1 OF



TOFAP.ADC TOFAP.UTT 10-Sep-75 21:41 UTT #:. 10 PAGE 1 OF 2

PHRASE 10 CAN YOU GIVE ME EVER ING ABOUT GANYMEDE (243600/249400)



S2: Segmentation -- Hand Corrected Machine Segmentation

DE GANYMEDE GANYMEDE GANYMEDE GANYMEDE GANYMEDE GANYMEDE GANYMEDE GANYMEDE GANYMEDE GANYMEDE GAN

258 262 266 270 274 278 282 286 290 294

298 302 306 310 314 318 322 326 330 334

338 342 346 350 354 358 362 368 370 374

378 382 386 390 394 398 402 406 410 414

418 422 426 430 434 438 442 446 450 454

L1: Labeling Evaluations

The following are evaluations of labeling for a number of the parametric representations, distance metrics, and labeling target sets investigated. The entire design space for labeling could not be covered in this appendix. The first table in each case contains two different acceptance criteria for templates: position (POS) and relative distance (RELDST). This latter is the difference between the score of the best template and that of the template in question. For either criterion, the target class accuracy (CL-SCOR) and the target class branching factor (CL-POS and BRNCH) are given for a range of acceptance values. The second display is the confusion matrix for the evaluation run. Entries are conditional probabilities * 100.

BF vs. Accuracy and Confusion Matrix
Number of clusters: 63
Distance Metric: EUC
Parametric Representation: SPG
Data File: TAP (speaker - CC, task - News)

Total Count: 1410

POS	CL-POS	CL-%COR	RELDST	CL-%COR	BRNCH
1	1.00	24.58	.00	24.86	1.07
2	1.93	42.44	1.00	30.01	1.42
3	2.84	54.03	2.00	36.30	1.83
4	3.72	60.10	3.00	43.22	2.31
5	4.57	64.83	4.00	50.78	2.90
6	5.40	68.71	5.00	56.00	3.54
7	6.19	71.54	6.00	60.52	4.20
8	6.97	74.15	7.00	64.76	4.86
9	7.74	76.41	8.00	68.36	5.54
10	8.45	78.60	9.00	71.75	6.22
11	9.18	79.73	10.00	74.36	6.92
12	9.88	81.14	11.00	75.78	7.55
13	10.60	83.05	12.00	77.05	8.20
14	11.33	84.53	13.00	78.39	8.81
15	12.02	85.31	14.00	79.86	9.41
16	12.71	85.95	15.00	80.86	10.04
17	13.39	86.58	16.00	81.85	10.66
18	14.08	87.01	17.00	82.98	11.23
19	14.72	87.64	18.00	83.97	11.83
20	15.34	88.35	19.00	84.89	12.39
21	15.94	88.63	20.00	85.52	12.97

	-	B	D	G	P	T	K	F	TH	V	DH	S	SH	Z	ZH	HH	DX	M	N	NX	EN	W	R	L	Y	EL	UW	UH	OW	AO	AA	AX	ER	EH	EY	AE	IX	IH	IY
-	71	7	1	3	5	.	1	.	.	1	1	.	.	.	.	.	.	5	.	.	1	.	.	.	4	.	.	.	.	.	.	.	.	.	.	.	.	1	.
B	5	31	7	2	14	.	.	.	.	.	5	.	.	.	.	.	.	33	.	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.	.
D	5	11	16	2	9	5	.	.	4	2	.	.	2	.	.	2	5	11	2	4	.	.	.	.	16	.	.	.	.	.	.	2	.	.	.	4	.	2	.
G	2	9	16	2	9	.	21	.	.	2	.	.	7	.	2	.	.	16	2	.	.	.	2	.	7	.	.	.	.	.	.	.	.	.	.	.	.	.	.
P	25	.	.	6	60	.	.	.	.	6	.	.	.	.	.	.	.	.	.	.	6	.	.	.	6	.	.	.	.	.	.	.	.	.	.	.	.	.	.
T	8	.	2	.	10	20	8	8	.	4	.	.	10	.	.	4	4	2	.	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	2	.
K	.	12	.	.	.	3	33	3	3	.	.	.	27	3	.	3	6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	6	.	.	.	.	.	.	.
F	47	.	.	16	16	.	.	5	.	.	.	.	6	.	.	.	.	11	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TH	25	6	.	10	25	.	.	16	6	.	.	.	.	.	.	16	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
V	3	.	7	3	.	.	.	.	.	3	3	.	.	.	.	.	30	7	.	.	.	10	.	.	23	.	.	.	.	.	.	.	3	.	.	.	3	3	.
DH	21	7	21	7	7	.	.	.	.	.	7	.	.	.	.	.	.	14	.	7	.	7	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
G	.	2	.	.	3	6	25	2	.	.	.	25	20	12	.	2	2	.	.	.	.	.	2	.	2	.	.	.	.	.	.	.	.	.	.	.	2	.	.
SH	.	.	.	.	.	.	67	.	.	.	.	33	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Z	.	.	.	.	3	6	12	.	.	.	.	20	16	26	.	.	3	3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3	.	.
ZH	.	.	.	.	.	.	75	.	.	.	.	.	25	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
HH	.	.	.	.	.	.	.	.	.	.	.	.	17	.	.	17	17	.	.	.	.	.	.	.	50	.	.	.	.	.	.	.	.	.	.	.	.	.	.
DX	.	.	10	.	.	.	.	.	.	.	.	.	.	.	.	.	50	.	.	.	.	.	.	.	30	.	.	.	.	.	.	.	.	.	.	.	.	10	.
M	.	4	11	.	.	.	.	.	.	.	4	.	.	.	.	.	4	21	21	21	14	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
N	.	1	30	.	.	.	1	.	.	.	.	.	.	.	.	.	9	11	11	32	2	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.
NX	.	.	11	.	.	.	.	.	.	.	.	.	.	.	.	.	.	5	.	74	.	.	.	.	.	.	.	.	.	.	.	11	.	.	.	.	.	.	.
EN	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
W	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	47	7	33	.	.	.	.	13	.	.	.	.	.	.	.	.	.	.	.	.	.
R	.	.	.	2	.	.	.	.	.	.	.	.	2	.	.	.	2	2	9	2	.	7	16	2	2	.	4	.	.	2	2	.	38	.	.	4	.	2	.
L	2	2	.	6	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.	11	.	7	2	26	2	13	.	.	.	6	4	6	.	6	2	.	6	.	.
Y	.	.	.	.	.	.	.	.	.	.	.	.	13	.	.	.	.	6	.	.	.	.	.	.	38	.	.	.	.	.	.	.	.	.	.	26	.	6	13
EL	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.100	.	.	.	.	.	.	.	.	.	.	.	.	.	.
UW	3	.	.	.	.	.	.	.	.	.	.	.	.	3	.	.	.	.	.	6	.	.	.	.	8	.	5	.	.	.	.	5	.	.	26	36	.	8	3
UH	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	8	.	.	.	.	.	8	8	.	.	8	16	.	.	.	.	23	.
OW	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	8	8	.	23	.	.	8	15	.	15	.	.	.	.	.	23	.
AO	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3	.	.	.	23	7	10	.	30	.	.	.	7	.	3	7	10	.	.	.	.	.	.
AA	.	4	.	4	.	.	.	.	.	4	.	.	.	.	.	.	.	.	.	4	.	.	4	.	.	.	.	4	.	12	4	30	.	.	.	19	.	4	.
AX	.	2	2	2	2	2	1	.	.	2	.	.	.	1	.	.	2	.	.	2	.	.	.	.	3	2	2	6	1	7	5	31	2	4	.	10	1	9	2
ER	.	4	.	8	.	.	.	.	.	4	.	.	.	.	.	.	.	.	17	.	.	.	4	.	4	.	4	.	.	.	.	.	33	.	.	13	.	4	4
EH	.	.	.	.	1	1	.	.	.	.	.	.	.	.	.	1	1	1	.	.	.	.	1	.	6	.	.	.	.	.	11	4	1	14	.	28	.	26	2
EY	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	20	.	.	.	.	.	.	.	.	.	.	80	.	.	.
AE	.	.	.	3	3	.	6	.	.	.	.	.	.	.	.	.	6	.	3	.	.	.	6	.	3	.	.	3	.	3	.	.	0	3	3	42	.	11	.
IX	.	.	6	6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	6	.	.	.	.	.	25	.	.	.	.	.	.	19	6	.	.	6	6	6	13
IH	.	.	1	.	.	.	1	.	.	.	.	.	.	1	.	.	1	.	1	.	.	.	.	.	20	.	.	.	.	.	.	2	1	6	5	32	.	21	6
IY	.	1	1	2	1	.	.	.	.	.	.	.	.	.	.	1	2	1	.	.	.	1	.	.	27	.	.	.	.	.	.	6	.	.	2	16	.	22	16

BF vs. Accuracy and Confusion Matrix
Number of clusters: 63
Distance Metric: COP
Parametric Representation: SPG
Data File: TAP (speaker - CC, task - News)

Total Count: 1416

POS	CL-POS	CL-%COR	PELDST	CL-%COR	BRNCH
1	1.00	23.73	.00	24.01	1.07
2	1.93	41.24	1.00	31.57	1.55
3	2.86	54.94	2.00	41.10	2.11
4	3.75	61.94	3.00	49.79	2.80
5	4.60	66.74	4.00	57.34	3.53
6	5.45	70.62	5.00	62.85	4.25
7	6.27	73.73	6.00	67.44	4.97
8	7.05	75.56	7.00	70.97	5.65
9	7.82	77.26	8.00	72.88	6.32
10	8.56	78.95	9.00	75.14	6.99
11	9.28	80.86	10.00	77.40	7.67
12	9.97	81.71	11.00	79.03	8.31
13	10.69	83.40	12.00	79.87	8.88
14	11.42	84.46	13.00	80.65	9.42
15	12.10	85.52	14.00	81.71	9.97
16	12.78	86.23	15.00	82.56	10.49
17	13.47	87.15	16.00	83.47	11.01
18	14.15	87.43	17.00	83.69	11.50
19	14.79	87.99	18.00	84.25	11.97
20	15.42	88.35	19.00	84.89	12.42
21	16.03	88.63	20.00	85.45	12.84

	-	B	D	G	P	T	K	F	TH	V	DH	S	SH	Z	ZH	HH	DX	M	N	NX	EN	W	R	L	Y	EL	UW	UH	OW	AO	AA	AX	ER	EH	EY	AE	IX	IH	IY
-	70	6	2	2	4	.	1	1	.	1	1	.	.	.	.	.	2	6	.	.	1	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	1	.
B	5	21	5	2	14	.	.	.	.	.	2	.	.	.	.	.	.	40	.	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.	.
D	5	7	14	2	9	4	4	.	2	2	.	.	2	.	.	.	9	12	.	4	.	.	.	.	.	21	.	.	.	.	.	.	.	.	4	.	2	.
G	2	9	16	2	7	.	23	.	.	.	.	.	7	.	2	.	.	19	2	.	.	.	2	.	7	.	.	.	.	.	.	.	.	.	.	.	.	.
P	25	.	.	6	44	.	6	.	.	6	.	.	.	.	.	.	.	.	.	6	.	.	.	.	6	.	.	.	.	.	.	.	.	.	.	.	.	.
T	6	.	2	.	14	14	16	6	.	4	.	.	22	.	.	4	6	2	.	.	.	.	2	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.
K	.	15	.	.	.	6	39	3	3	.	.	.	21	.	.	.	6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3	.	.	.	.	3	.
F	47	.	.	16	11	5	.	.	.	.	.	.	11	.	.	.	.	11	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TH	30	15	.	5	15	5	.	.	5	.	.	.	10	.	.	15	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
V	3	.	3	3	.	.	.	.	.	3	.	.	.	.	.	.	20	10	.	3	.	20	.	.	23	.	.	.	.	.	.	.	.	3	.	3	.	.	3
DH	21	7	21	.	7	.	.	.	.	.	7	.	.	.	.	.	.	14	7	7	.	.	.	7	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
S	2	.	.	.	3	5	32	3	.	.	.	28	12	11	.	.	.	.	2	.	.	.	2	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.
SH	.	.	.	.	.	.	100	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Z	.	.	3	.	3	.	26	.	.	.	.	24	9	29	.	.	.	.	3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3	.	.
ZH	.	.	.	.	.	.	100	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
HH	.	.	.	.	.	.	.	.	.	.	.	.	17	.	.	17	17	.	.	.	.	.	.	.	.	50	.	.	.	.	.	.	.	.	.	.	.	.	.
DX	.	.	10	.	.	.	.	.	.	.	.	.	.	.	.	.	60	.	.	.	.	.	.	.	20	.	.	.	.	.	.	.	.	.	.	.	10	.
M	.	4	11	.	.	.	.	.	.	.	4	.	.	.	.	.	7	10	10	29	11	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
N	.	1	27	.	.	.	.	.	.	.	.	.	.	.	.	.	7	11	9	36	5	.	.	.	2	.	.	.	.	.	1	.	.	.	.	.	.	.	.
NX	.	.	11	.	.	.	.	.	.	.	.	.	.	.	.	.	.	5	.	74	.	.	.	.	.	.	.	.	.	.	.	11	.	.	.	.	.	.	.
EN	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
W	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	53	7	20	.	.	.	20	.	.	.	.	.	.	.	.	.	.	.	.	.	.
R	.	.	.	.	2	.	.	.	.	.	.	.	2	.	.	.	4	2	9	2	.	7	18	.	2	.	2	.	.	2	2	.	40	.	.	4	.	.	.
L	2	2	.	6	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.	7	.	7	2	26	.	20	.	.	.	6	4	11	.	4	2	.	.	.	.
Y	.	.	.	.	.	.	6	.	.	.	.	.	.	.	.	.	.	6	.	.	.	.	.	.	63	.	.	.	.	.	.	.	.	.	.	19	.	.	6
EL	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	100	.	.	.	.	.	.	.	.	.	.	.	.	.	.
UW	3	.	.	.	.	.	.	.	.	.	.	.	.	3	.	.	.	.	.	5	.	.	.	.	15	5	5	.	.	.	.	3	.	.	15	36	.	10	.
UH	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	.	.	0	0	0	.	.	0	30	0	.	.	.	.	.	15	.
OW	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	9	0	23	.	.	.	23	.	15	0	.	.	.	.	.	15	.
AO	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	23	7	17	30	.	.	.	3	.	7	3	7	.	3	.	.	.	.
AA	.	.	.	.	4	.	.	.	.	4	.	.	.	.	.	.	4	.	.	.	.	.	4	.	.	.	4	.	12	12	42	.	.	.	12	.	4	.
AX	.	1	2	.	2	2	1	.	.	1	.	.	.	1	.	.	7	1	.	2	.	1	2	.	3	2	2	4	2	5	6	30	4	2	.	12	1	0	3
ER	.	4	.	4	.	.	.	.	.	.	4	.	.	.	.	.	.	.	17	.	.	.	.	.	4	.	.	.	.	.	.	4	46	.	.	13	.	.	4
EH	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	2	.	.	.	.	.	1	.	7	.	.	.	.	.	10	1	4	12	.	37	.	21	2
EY	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	20	.	.	.	.	.	.	.	.	.	.	80	.	.	.
AE	.	.	3	.	6	.	3	.	.	.	.	.	.	.	.	.	6	.	.	.	.	.	6	3	3	.	.	3	.	3	.	0	3	6	44	.	6	.
IX	.	.	6	6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	6	.	.	.	.	31	.	.	.	.	.	19	6	.	.	13	6	6	.
IH	.	.	1	.	.	1	1	.	.	.	.	.	.	.	.	.	1	.	.	1	.	.	.	.	23	.	.	.	.	.	1	1	9	2	36	.	14	7
IY	.	.	.	2	1	.	.	.	.	.	.	.	.	.	.	.	2	2	.	.	.	.	1	.	41	.	.	.	.	.	6	.	.	1	17	.	15	11

BF vs. Accuracy and Confusion Matrix
Number of clusters: 63
Distance Metric: SIG
Parametric Representation: SPG
Data File: TAP (speaker - CC, task - News)

Total Count: POS	1416 CL-POS	CL-%COR	RELDST	CL-%COR	BRNCH
1	1.00	25.56	.00	25.85	1.87
2	1.93	42.58	1.00	31.26	1.41
3	2.84	53.88	2.00	38.06	1.82
4	3.70	60.52	3.00	44.07	2.20
5	4.53	65.75	4.00	50.07	2.76
6	5.33	69.49	5.00	56.07	3.30
7	6.12	72.81	6.00	61.23	3.88
8	6.87	75.35	7.00	64.83	4.48
9	7.61	77.12	8.00	68.95	5.04
10	8.37	78.60	9.00	70.55	5.66
11	9.07	80.08	10.00	73.09	6.26
12	9.78	81.64	11.00	75.00	6.87
13	10.50	83.40	12.00	76.41	7.45
14	11.28	84.60	13.00	77.75	8.02
15	11.89	85.45	14.00	79.24	8.63
16	12.59	85.95	15.00	80.01	9.20
17	13.26	86.72	16.00	81.00	9.77
18	13.93	87.36	17.00	81.92	10.31
19	14.57	88.21	18.00	82.84	10.86
20	15.19	88.49	19.00	83.62	11.36
21	15.77	88.63	20.00	84.68	11.85

	-	B	D	G	P	T	K	F	TH	V	DH	S	SH	Z	ZH	HH	DX	M	N	NX	EN	W	R	L	Y	EL	UW	UH	OW	AO	AA	AX	ER	EH	EY	AE	IX	IH	IY
-	70	7	2	5	5	.	1	1	.	1	1	.	.	.	.	.	.	2	1	.	.	.	.	.	3	.	.	.	.	.	.	1	.	.	.	.	.	.	1
B	12	36	10	2	14	.	.	.	.	.	12	.	.	.	.	.	.	12	.	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.	.
D	9	14	16	4	12	5	.	.	.	4	.	.	.	.	.	2	4	5	2	.	.	.	2	.	14	.	.	.	.	.	.	2	.	.	.	4	.	2	2
G	5	9	21	.	7	5	14	.	.	2	.	.	7	.	5	.	2	9	2	.	.	.	5	.	7	.	.	.	.	.	.	.	.	.	.	.	.	.	.
P	25	.	.	6	44	.	.	.	.	6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	13	.	.	.	.	.	.	6	.	.	.	.	.	.	.
T	6	2	4	.	16	20	6	4	.	4	.	.	18	.	2	6	4	.	.	.	.	.	.	.	4	.	.	.	.	.	.	.	.	2	.	2	.	.	.
K	.	12	.	.	3	.	33	3	.	.	.	.	27	3	3	3	6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	8	.	.	.	.	.	.	.
F	37	.	5	11	16	.	.	11	.	.	.	.	5	.	.	5	.	11	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TH	10	26	.	10	26	.	.	10	5	.	.	.	.	.	.	15	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
V	3	.	3	.	.	.	.	.	.	3	3	.	.	.	.	.	20	13	3	.	.	13	.	.	30	.	.	.	.	.	.	.	.	3	.	.	.	3	.
DH	14	14	29	7	7	.	.	.	.	.	7	.	.	.	.	.	.	7	7	.	.	.	.	.	.	.	.	.	.	.	.	7	.	.	.	.	.	.	.
S	.	2	.	.	6	2	22	2	.	.	.	31	18	9	.	2	2	.	.	.	.	.	.	.	5	.	.	.	.	.	.	.	.	.	.	.	2	.
SH	.	.	.	.	.	.	67	.	.	.	.	33	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Z	.	8	8	.	.	8	18	.	.	8	.	29	12	26	.	.	.	.	.	.	.	.	.	.	8	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ZH	.	.	.	.	.	.	25	.	.	.	.	.	50	.	.	.	.	.	.	.	.	.	.	.	25	.	.	.	.	.	.	.	.	.	.	.	.	.	.
HH	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	33	17	.	.	.	.	.	.	.	33	.	.	.	.	.	.	.	.	.	.	17	.	.	.
DX	.	.	10	.	.	.	.	.	.	.	.	.	.	.	.	.	40	.	.	.	.	.	.	.	40	.	.	.	.	.	.	.	.	10	.	.	.	.	.
M	4	.	14	.	.	.	.	.	.	14	.	.	.	.	.	.	4	36	11	18	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
N	.	1	37	.	.	.	1	.	.	.	.	.	.	.	.	.	7	15	13	23	.	.	.	.	1	.	.	.	.	.	.	1	.	.	.	.	.	.	.
NX	.	.	11	.	.	.	.	.	.	.	.	.	.	.	.	.	.	5	.	74	.	.	.	.	.	.	.	.	.	.	.	5	.	.	.	5	.	.	.
EN	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
W	.	.	13	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	7	7	.	67	.	.	7	.	.	.	.	.	.	.	.	.	.	.	.	.	.
R	.	.	.	.	2	.	.	.	.	.	.	.	2	.	.	.	2	4	4	2	.	11	13	.	2	.	2	.	.	2	2	.	44	2	.	.	.	2	.
L	2	2	.	4	2	.	.	.	.	.	.	.	.	.	.	.	2	.	4	6	.	7	2	26	.	11	.	.	.	6	2	11	.	6	2	2	4	2	.
Y	.	.	.	.	.	.	.	.	.	.	.	.	13	.	.	.	.	6	.	.	.	.	.	.	30	.	.	.	.	.	.	.	.	.	.	19	.	13	13
EL	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	100	.	.	.	.	.	.	.
UW	3	.	.	.	.	.	.	.	.	.	.	.	.	3	.	.	.	.	.	3	.	.	.	.	10	5	13	3	.	.	.	3	.	.	8	33	.	13	5
UH	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	8	.	.	.	.	.	8	.	.	16	38	15	.	.	.	15	.
OW	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	8	8	.	23	.	.	8	23	.	15	.	.	.	.	15	.
AO	.	.	.	.	.	.	.	.	.	3	.	.	.	.	.	.	3	.	.	.	.	23	7	17	.	17	.	.	.	3	.	7	7	10	.	.	3	.	.
AA	.	4	.	4	.	.	.	.	.	4	.	.	.	.	.	.	.	.	.	.	.	.	4	.	.	8	.	8	.	8	8	38	.	.	.	8	.	8	.
AX	1	1	2	.	2	.	1	.	.	2	.	.	.	1	.	.	1	2	.	1	.	.	.	1	3	2	1	8	1	9	3	30	2	2	.	7	2	16	2
ER	.	.	.	4	4	.	.	.	.	13	.	.	.	.	.	.	4	.	8	.	.	.	4	.	.	.	4	.	.	.	.	.	33	.	.	4	4	8	8
EH	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	2	.	7	.	.	.	.	1	9	5	1	15	.	31	.	25	1
EY	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	20	.	.	.	.	.	.	.	.	.	.	80	.	.	.
AE	.	.	3	.	3	.	8	.	.	3	.	.	.	.	.	.	6	.	.	.	.	.	8	.	3	.	.	.	.	3	.	3	8	6	.	44	.	8	3
IX	.	.	6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	6	6	.	.	.	.	25	.	.	.	.	.	.	25	6	.	.	6	6	.	13
IH	.	.	1	.	.	.	1	.	.	.	.	.	.	.	.	.	1	.	.	1	.	.	.	.	26	.	.	.	.	1	.	1	1	6	.	38	.	17	6
IY	.	1	1	2	1	.	.	.	.	.	.	.	.	.	.	.	1	1	1	.	.	1	.	.	26	.	.	.	1	.	.	5	.	1	.	15	.	26	16

BF vs. Accuracy and Confusion Matrix
Number of clusters: 87
Distance Metric: ITK
Parametric Representation: ACS
Data File: TAP (speaker - CC, task - Nums)

Total Count: 1416

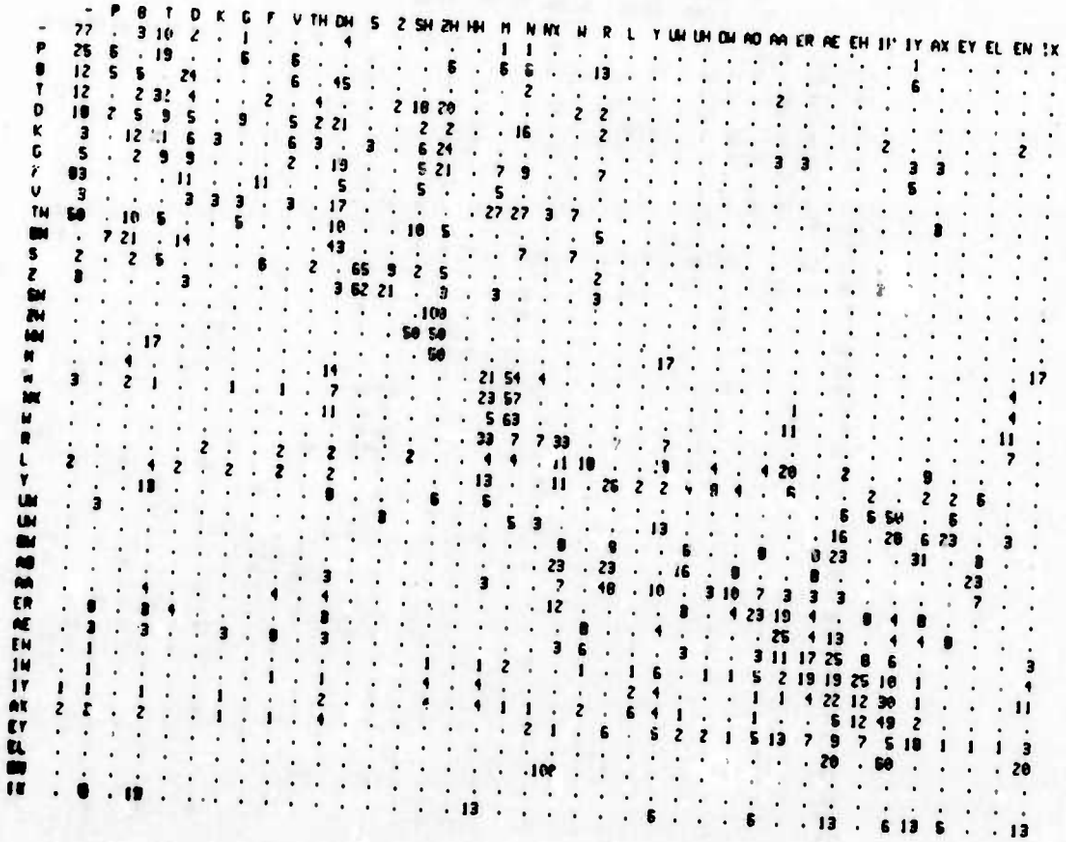
POS	CL-POS	CL-%COR	PELOST	CL-%COR	BRNCH
1	1.00	28.67	.00	29.38	1.16
2	1.86	44.42	1.00	32.06	1.52
3	2.69	54.59	2.00	34.53	1.74
4	3.50	59.96	3.00	36.51	1.90
5	4.29	64.12	4.00	38.35	2.05
6	5.06	67.02	5.00	40.02	2.19
7	5.81	70.13	6.00	43.00	2.36
8	6.55	72.32	7.00	45.27	2.63
9	7.24	74.72	8.00	47.32	2.80
10	7.91	76.98	9.00	48.52	2.97
11	8.55	78.74	10.00	50.20	3.17
12	9.18	79.52	11.00	52.05	3.45
13	9.78	80.79	12.00	53.46	3.68
14	10.36	81.57	13.00	55.16	3.89
15	10.93	82.70	14.00	56.57	4.13
16	11.45	83.55	15.00	58.05	4.44
17	11.97	84.75	16.00	59.75	4.67
18	12.51	85.52	17.00	60.88	4.94
19	13.04	86.86	18.00	62.08	5.18
20	13.55	87.43	19.00	63.49	5.39
21	14.05	88.28	20.00	64.34	5.67

	-	B	D	G	P	T	K	F	TH	V	DH	S	SH	Z	ZH	HH	DX	M	N	NX	EN	W	R	L	Y	EL	UW	UH	OW	AO	AA	AX	ER	EH	EY	AE	IX	IH	IY
-	74	1	5	2	2	.	1	1	3	.	.	.	2	.	.	.	.	4	2	.	.	.	.	.	2	.	.	.	.	.	.	1	.	.	.	.	.	1	1
B	.	17	52	.	10	.	.	.	.	2	.	.	.	.	.	.	2	10	5	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	.
D	14	7	12	2	4	.	5	2	.	2	2	.	7	.	4	.	9	2	11	.	.	.	.	.	14	.	.	.	.	.	.	.	.	.	.	.	.	4	2
G	5	14	16	.	2	.	9	.	.	2	2	.	9	.	14	.	.	.	14	.	.	2	.	.	5	.	.	.	.	.	.	.	.	.	.	.	2	.	2
P	19	.	.	6	19	.	.	.	6	.	.	.	6	.	13	.	13	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	13	6
T	10	.	2	.	2	2	2	10	2	2	.	.	12	2	26	2	8	.	4	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	2	2	.
K	3	6	.	.	6	.	.	.	.	.	.	.	36	.	24	.	.	.	.	.	.	3	9	.	3	.	.	3	.	.	.	3	.	.	.	.	3	.	.
F	50	.	.	5	5	5	.	.	.	.	.	.	5	.	.	.	.	11	5	.	.	.	.	6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TH	30	.	.	10	10	.	.	10	10	5	.	.	25	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
V	.	3	7	10	3	.	.	.	.	23	3	.	.	.	.	.	10	3	3	.	.	23	.	.	.	.	.	.	.	.	.	3	.	.	.	.	3	3
DH	21	.	29	.	.	.	.	.	.	.	.	.	.	.	.	.	7	7	29	.	.	.	.	.	.	.	.	.	.	.	7	.	.	.	.	.	.	.	.
S	5	2	.	.	2	2	.	2	2	.	.	42	12	15	10	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
SH	.	.	.	.	.	.	.	.	.	.	.	.	33	.	67	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Z	3	.	.	3	3	.	.	3	.	3	.	41	3	32	3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3	3
ZH	.	.	.	.	.	.	.	.	.	.	.	.	50	.	25	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	25	.
HH	.	.	.	.	.	.	.	.	.	.	.	.	17	.	.	.	17	.	.	.	.	.	.	.	33	.	.	.	.	.	.	.	.	.	.	.	33	.
DX	.	.	10	.	10	.	.	.	.	.	.	.	10	.	.	.	40	.	.	.	.	.	.	.	10	.	.	.	.	.	.	.	.	.	.	.	20	.
M	.	.	7	.	.	.	.	.	.	7	.	.	.	.	.	.	.	29	29	7	4	.	.	4	.	14	.	.	.	.	.	.	.	.	.	.	.	.
N	.	1	3	1	.	.	.	.	.	.	.	.	.	.	.	.	4	7	61	16	2	.	.	.	2	.	1	.	.	.	.	1	.	.	.	.	.	.
NX	.	.	.	.	.	.	.	.	.	5	.	.	.	.	.	.	.	.	42	21	.	.	.	21	.	5	.	.	.	.	.	.	.	.	.	.	6	.	.
EN	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
W	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	20	.	60	.	7	13	.	.	.	.	.	.	.	.	.	.	.	.
P	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	7	.	2	.	.	4	29	.	.	.	11	.	.	2	.	16	9	7	.	2	.	2	7
L	2	4	.	2	.	.	.	.	.	2	.	.	.	.	.	.	.	4	.	2	2	6	.	26	.	15	.	.	17	9	.	9	.	.	.	.	.	2	.
Y	.	.	.	.	.	.	6	.	.	.	.	6	6	.	.	.	.	.	.	.	.	.	.	.	31	.	.	.	.	.	.	.	.	.	.	.	.	25	25
EL	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	100	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
UW	.	.	.	.	.	.	.	.	.	.	.	.	.	3	.	.	.	.	.	8	.	.	.	3	.	.	10	.	.	.	.	3	.	5	.	.	21	41
UH	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	31	8	23	.	.	.	15	.	.	.	23	.
OW	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	15	8	.	8	.	.	31	8	8	8	.	.	.	.	16	.
AO	.	3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3	.	.	.	.	7	13	17	.	10	17	.	.	20	.	10	.	.	.	.	.	.	.
AA	.	4	.	.	.	.	.	.	.	8	.	.	.	.	.	.	4	.	.	.	.	.	.	15	.	12	.	8	.	4	4	12	.	.	.	4	8	19	.
AX	1	2	.	3	1	.	1	1	.	2	.	.	.	.	.	.	1	.	2	2	.	.	1	1	1	1	4	8	1	4	2	27	1	3	1	6	3	16	5
ER	.	.	.	.	4	.	.	.	.	.	.	.	.	.	.	.	8	.	8	.	.	17	.	4	.	8	.	.	.	.	4	29	4	.	.	.	.	.	13
EH	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	.	.	.	.	.	.	.	6	.	.	1	.	5	9	1	1	32	.	2	2	37	.
EY	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	80	.	.	.	.	.	.	.	.	.	.	.	.	20	.
AE	.	.	.	.	3	.	3	.	3	.	.	.	.	.	.	.	11	.	.	.	.	.	3	.	3	.	.	6	.	.	.	3	3	3	3	11	.	39	8
IX	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	13	6	6	.	.	.	6	6	.	.	.	.	.	.	13	.	.	.	6	.	31	13
IH	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	30	.	.	.	1	.	.	2	.	15	.	.	4	37	9
IY	.	.	.	.	1	.	2	.	.	2	.	1	.	.	.	.	1	.	.	.	.	.	.	1	41	.	.	.	1	.	1	5	.	1	.	1	2	16	22

BF vs. Accuracy and Confusion Matrix
Number of clusters: 76
Distance Metric: EUC
Parametric Representation: ASR
Data File: TAP (speaker - CC. test - News)

Total Count: 1446

POS	CL-POS	CL-%COR	PELOST	CL-%COR	3PNCH
1	1.00	27.10	.00	27.60	1.34
2	1.94	39.12	1.00	32.36	1.65
3	2.87	50.43	2.00	35.21	1.95
4	3.73	58.10	3.00	39.05	2.33
5	4.56	63.09	4.00	42.82	2.69
6	5.39	68.71	5.00	46.87	3.10
7	6.17	71.83	6.00	50.71	3.53
8	6.83	74.18	7.00	55.05	4.00
9	7.49	76.89	8.00	59.17	4.47
10	8.17	77.81	9.00	61.74	4.94
11	8.80	79.30	10.00	64.58	5.49
12	9.59	81.01	11.00	67.43	6.07
13	10.27	82.15	12.00	69.42	6.59
14	10.88	82.79	13.00	71.76	7.12
15	11.54	84.07	14.00	73.76	7.70
16	12.16	84.64	15.00	74.89	8.27
17	12.76	85.42	16.00	77.03	8.81
18	13.42	86.27	17.00	78.45	9.36
19	14.07	87.27	18.00	79.45	9.93
20	14.64	88.41	19.00	80.37	10.46
21	15.23	88.98	20.00	81.29	11.02



BF vs. Accuracy and Confusion Matrix
Number of clusters: 76
Distance Metric: COR
Parametric Representation: ASH
Data File: TAP (speaker = CC, task = Navi)

Total Count: 1406

POS	CL-POS	CL-%COR	PELOST	CL-%COR	BPNCH
1	1.00	24.96	.00	26.24	1.45
2	1.92	36.98	1.00	34.50	2.06
3	2.80	48.29	2.00	42.46	2.75
4	3.64	56.26	3.00	49.79	3.49
5	4.38	62.38	4.00	54.77	4.19
6	5.11	66.93	5.00	59.32	5.00
7	5.83	70.34	6.00	62.52	5.70
8	6.43	72.48	7.00	65.58	6.36
9	7.02	74.40	8.00	67.50	7.00
10	7.59	76.03	9.00	68.61	7.50
11	8.21	77.95	10.00	70.20	8.16
12	8.84	79.87	11.00	71.91	8.76
13	9.44	81.58	12.00	73.51	9.32
14	9.98	82.93	13.00	74.82	9.83
15	10.55	84.50	14.00	75.89	10.41
16	11.12	86.20	15.00	77.38	10.99
17	11.64	86.91	16.00	78.38	11.50
18	12.23	87.62	17.00	78.88	12.03
19	12.81	88.12	18.00	80.16	12.55
20	13.33	89.26	19.00	81.29	13.02
21	13.89	89.69	20.00	82.01	13.51

	-	P	B	T	D	K	G	F	V	TH	DH	S	Z	SH	ZH	HH	M	N	NX	W	P	L	Y	UW	UH	OW	AO	AA	ER	AE	EH	IH	IY	AX	EY	EL	EN	IX
-	49	1	6	14	17	.	3	2	.	1	3	.	.	.	1	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	1	1	.	.	.	.	.
P	6	6	.	19	13	.	.	.	6	.	.	.	13	6	.	.	6	.	.	6	.	6	.	6	.	.	.	.	.	.	.	.	13	.	.	.	.	.
B	2	2	.	.	26	5	.	.	.	.	52	.	.	.	.	.	5	.	.	2	.	.	.	.	.	.	2	.	.	.	.	.	2	.	.	.	.	.
T	6	2	.	26	4	.	2	2	.	2	.	4	.	28	10	.	.	2	.	.	4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
D	5	2	4	11	11	2	4	2	2	2	16	.	.	5	5	.	2	10	.	.	.	2	2	.	.	.	.	.	.	.	.	.	5	.	.	.	.	.
K	.	.	.	3	6	3	.	3	9	3	.	.	3	12	45	.	.	.	.	.	.	.	.	.	6	.	.	.	3	.	.	.	3	.	.	.	.	.
G	2	2	2	9	9	.	.	.	2	.	21	.	.	.	20	.	.	12	.	.	7	.	.	.	.	.	.	.	.	.	.	.	6	.	.	.	.	.
F	37	.	.	.	11	.	.	5	.	1.	11	5	.	11	.	5	5	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
V	.	.	.	.	3	3	3	.	.	.	23	.	.	.	.	.	20	27	.	10	3	.	.	3	.	.	.	.	.	3	.	.	.	.	.	.	.	.
TH	10	.	.	15	5	.	5	5	.	.	5	15	.	30	5	5	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
DH	.	.	7	.	7	.	.	.	.	.	50	.	.	.	14	.	.	14	.	.	.	7	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
S	.	.	.	5	.	.	.	2	.	2	.	66	8	6	6	.	.	2	.	.	2	.	.	.	.	.	.	.	.	.	.	3	.	.	.	.	.	.
Z	.	.	.	3	3	.	.	.	.	.	.	68	15	.	6	.	.	6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
SH	.	.	.	.	.	.	.	.	.	.	.	.	.	.100	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ZH	.	.	.	.	.	.	.	.	.	.	.	.	25	75	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
HH	.	.	.	67	.	.	.	.	.	.	.	.	.	17	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	17	.	.	.	.	.	.	.
M	.	.	.	.	11	.	.	.	.	.	14	.	.	.	.	.	14	54	4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	4	.
N	1	.	.	1	13	.	.	2	.	.	13	.	.	1	.	.	6	55	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	6	.
NX	.	.	.	.	11	.	.	.	.	.	16	.	.	.	.	.	16	47	.	.	.	.	.	.	.	.	.	.	11	.	.	.	.	.	.	.	.	.
W	.	.	.	.	13	.	.	.	.	.	7	.	.	.	.	.	13	13	.	33	.	.	13	7	.	.	.	.	.	.	.	.	.	.	.	.	.	.
R	.	.	.	.	.	4	.	.	.	.	2	.	.	2	.	.	4	2	.	7	10	2	.	13	.	4	2	2	22	.	.	2	2	9	.	.	.	.
I	.	.	.	4	4	.	.	.	2	.	2	.	2	.	.	.	2	.	.	4	.	37	.	9	2	13	6	.	.	.	2	2	2	4	.	6	.	.
Y	.	.	.	6	.	.	.	.	.	.	13	.	.	6	6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	13	.	56	.	.	.	.	.
UW	.	.	.	.	.	.	.	.	.	.	.	.	3	.	.	.	.	8	3	.	.	.	.	21	.	.	.	.	.	3	.	13	.	31	8	3	5	5	.
UH	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	8	8	8	.	.	.	.	.	.	8	.	15	15	.	.	38	.	.	.	.
OW	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	15	.	33	.	.	.	.	8	.	.	8	.	.	.	8	.	23	.	.	.
AO	.	.	.	.	.	.	.	.	.	.	3	.	.	.	.	.	3	.	.	3	.	27	.	3	.	7	13	.	3	.	13	.	.	.	23	.	.	.	.
AA	.	.	.	4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	8	.	12	.	.	8	12	15	15	4	4	4	.	8	8	.	.	.	.	.
ER	.	.	.	4	.	.	4	.	4	.	.	.	.	.	4	.	4	8	.	.	4	.	.	13	.	.	4	.	33	.	8	.	.	4	4	.	.	.	.
AE	.	.	.	3	.	.	3	.	3	.	3	.	.	.	.	.	.	.	.	.	6	.	.	.	6	.	.	3	8	17	31	6	11	3	.	.	.	.	.
EH	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	4	1	.	.	.	1	1	7	.	27	33	11	9	.	.	.	.	4
IH	.	.	.	.	.	.	.	.	.	.	1	.	.	1	1	.	2	.	.	1	.	.	4	1	1	.	.	.	1	7	30	5	26	1	.	.	1	6
IY	.	.	.	1	1	.	.	.	.	.	5	.	.	1	5	.	.	.	.	.	2	6	6	.	.	.	.	.	.	2	10	6	48	6	.	.	.	.
AX	.	1	.	1	.	1	.	.	1	.	3	.	.	1	.	.	.	1	1	.	1	8	.	5	7	1	2	5	7	9	13	10.	5	16	.	2	1	.
EY	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	40	.	40	.	.	.	.	20
EL	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.100	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
EN	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
IX	.	.	.	.	.	.	.	.	.	.	13	.	.	13	6	.	.	.	.	.	.	.	6	.	6	.	6	.	.	19	.	.	25	.	.	.	6	.

BF vs. Accuracy and Confusion Matrix
Number of clusters: 76
Distance Metric: LIK
Parametric Representation: ASA
Data File: TAP (speaker - CC, task - News)

Total Count: 1406

PDS	CL-PDS	CL-%COR	PELOST	CL-%COR	BRNCH
1	1.00	25.11	.00	26.17	1.16
2	1.63	35.49	1.00	44.45	3.11
3	2.30	44.10	2.00	61.10	5.42
4	2.94	50.36	3.00	70.77	7.57
5	3.53	55.33	4.00	77.67	10.26
6	4.12	59.17	5.00	83.07	12.30
7	4.69	61.74	6.00	86.49	14.67
8	5.24	64.44	7.00	89.69	16.37
9	5.75	66.71	8.00	91.68	17.77
10	6.25	69.49	9.00	92.67	18.89
11	6.74	72.33	10.00	93.46	20.05
12	7.15	74.25	11.00	94.31	20.95
13	7.67	76.24	12.00	94.59	21.72
14	8.11	78.02	13.00	95.02	22.63
15	8.64	79.09	14.00	95.52	23.50
16	9.14	80.58	15.00	95.80	24.12
17	9.52	81.85	16.00	95.87	24.70
18	9.92	82.86	17.00	96.16	25.42
19	10.32	83.64	18.00	96.44	25.83
20	10.70	84.35	19.00	96.58	26.26
21	11.07	86.21	20.00	96.94	26.66

	-	P	B	T	D	K	G	F	V	TH	DH	S	Z	SH	ZH	HH	M	N	NX	W	R	L	Y	UW	UH	OW	AO	AA	ER	AE	EH	IH	IY	AX	EY	EL	EN	IX
-	85	.	.	2	1	.	.	.	.	.	2	.	.	.	.	.	.	1	1	.	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.	5	.	.	.
P	31	6	.	31	.	6	6	.	.	.	.	.	.	.	.	.	.	6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	6	6	.	.	.	.	.
B	14	.	.	2	17	.	.	.	.	.	17	.	.	.	.	.	.	.	2	.	2	.	.	.	.	.	.	.	.	.	2	.	6	.	.	30	.	.	.
T	14	2	2	32	2	4	16	0	.	2	.	12	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
D	23	.	2	11	9	2	7	.	.	.	14	.	.	.	.	.	.	14	2	2	.	.	.	.	.	.	.	.	.	4	.	.	6	.	.	7	.	.	.
K	9	.	3	10	.	9	21	3	.	.	.	21	.	.	.	.	.	.	9	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	12	.	.	.
G	9	.	.	9	9	6	16	2	2	.	12	5	2	.	2	.	.	9	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	5	.	9	.	.	.
F	47	.	.	11	.	.	.	11	.	.	11	5	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	16	.	.	.
V	7	3	3	.	3	.	.	.	3	.	20	.	.	.	.	.	.	13	7	10	13	3	.	.	.	.	.	.	.	.	.	.	7	.	.	7	.	.	.
TH	15	.	5	16	.	.	10	20	.	20	6	10	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
DH	7	7	.	.	7	.	.	.	.	7	14	.	.	.	.	.	.	14	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
S	.	.	.	8	.	.	6	2	.	.	.	63	2	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	43	.	.	.
Z	3	.	.	.	.	3	.	6	.	.	.	41	12	.	.	.	.	.	3	.	.	.	.	.	.	.	.	.	.	.	.	.	2	3	.	.	14	.	.	.
SH	.	.	.	.	.	.	67	.	.	.	.	33	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3	29	.	.	.
ZH	.	.	.	.	.	.	100	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
HH	.	.	.	83	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
M	.	.	4	.	11	.	.	.	.	.	14	.	4	.	.	.	14	36	11	4	.	.	.	.	.	.	.	.	.	.	.	.	17	.	.	.	.	.	.
N	3	.	.	.	10	.	.	1	.	.	11	.	.	.	.	.	1	53	12	1	.	.	.	.	.	.	.	.	.	1	.	.	4	.	.	.	.	.	.
NX	.	.	.	.	16	11	.	.	.	.	5	.	.	.	.	.	5	47	16	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
W	.	.	.	.	7	.	.	.	.	.	.	.	.	.	.	.	7	27	13	40	.	7	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
R	2	7	.	2	.	.	.	7	.	.	4	.	.	.	.	.	.	2	2	11	13	.	.	.	.	.	.	.	4	9	2	2	31	.	.	.	.	.	.
L	7	26	.	.	.	.	.	.	2	.	4	.	.	.	.	.	2	2	8	6	2	22	.	2	.	.	.	11	.	8	.	4	.	.	.	.	.	.	.
Y	.	.	.	.	.	.	8	8	.	.	19	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	63	8	.	.	.	.	8
UW	.	.	.	.	3	.	3	.	.	.	.	.	3	.	.	.	.	.	.	3	.	3	6	6	.	.	.	.	.	.	.	.	62	5	10	.	.	.	.
UH	.	8	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	8	.	8	.	.	.	.	.	23	.	15	.	31	.	.	8	.	.	.	.
OW	.	23	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	23	6	15	.	.	.	.	.	15	.	.	8	.	.	8	.	.	.	.	.
AO	.	3	.	.	.	.	.	.	.	.	3	.	.	.	.	.	3	.	.	27	3	17	.	3	.	3	3	3	7	10	7	.	.	3	.	3	.	.	.
AA	4	8	.	.	.	.	.	.	.	.	.	.	.	.	.	.	4	.	.	12	.	12	.	.	.	.	.	18	4	12	.	19	.	.	4	.	.	4	.
ER	4	4	4	.	.	.	.	4	.	.	17	.	.	.	.	.	.	8	.	.	.	.	.	.	.	.	.	.	25	4	4	25	.	.	.	.	.	.	.
AE	.	3	.	6	.	3	3	.	.	.	3	.	.	.	.	.	.	3	.	.	.	.	.	.	.	.	.	.	.	.	42	8	22	3	.	6	.	.	.	.
EH	.	17	.	1	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	1	.	.	.	.	1	1	41	2	25	1	.	8	.	.	.	.
IH	1	7	.	2	.	1	4	.	.	.	1	1	.	.	.	.	.	.	1	1	.	1	1	1	.	.	.	.	.	8	4	48	18	.	6	.	.	1	.
IY	2	.	.	1	1	.	8	.	.	.	4	.	.	.	.	.	.	.	.	1	.	.	5	.	.	.	.	1	.	2	1	56	10	1	8	.	.	1	.
AX	2	8	.	2	2	2	.	2	.	.	2	.	.	.	.	.	1	1	2	2	1	2	6	.	1	1	1	.	7	.	19	6	25	2	4	2	.	.	.	.
EY	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	20	20	20	40	.	.	.	.	.	.
EL	100	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
EN	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
IX	.	.	.	8	8	.	6	6	.	.	6	.	.	.	.	.	.	19	.	6	.	6	.	.	.	.	.	.	.	.	.	8	19	.	6	6	.	.	6	.

DF vs. Accuracy and Confusion Matrix
Number of clusters: 76
Distance Metric: SIG
Parametric Representation: ASA
Data File: TAP (speaker - CC: task - News)

Total Count: 1406

POS	CL-POS	CL-%COR	PELOSI	CL-%COR	BRNCH
1	1.00	28.66	.00	29.30	1.33
2	1.90	41.25	1.00	35.70	1.79
3	2.80	50.57	2.00	41.68	2.28
4	3.84	56.54	3.00	47.51	2.88
5	4.47	61.88	4.00	51.71	3.42
6	5.20	66.15	5.00	54.62	3.98
7	5.96	69.77	6.00	58.39	4.56
8	6.62	72.62	7.00	61.52	5.11
9	7.22	74.04	8.00	64.79	5.69
10	7.83	75.82	9.00	67.50	6.30
11	8.51	77.81	10.00	70.20	6.89
12	9.15	80.01	11.00	72.05	7.50
13	9.74	81.37	12.00	74.11	8.09
14	10.30	82.57	13.00	75.60	8.62
15	10.85	83.57	14.00	77.45	9.16
16	11.37	84.42	15.00	79.02	9.72
17	11.87	85.46	16.00	80.09	10.21
18	12.43	85.85	17.00	80.51	10.71
19	12.99	86.49	18.00	81.51	11.15
20	13.50	87.62	19.00	82.01	11.55
21	14.04	88.05	20.00	83.07	11.93

	-	P	B	T	D	K	G	F	V	TH	DH	S	Z	SH	ZH	HH	M	N	NX	W	R	L	Y	UW	UH	OW	AO	AA	ER	AE	EH	IH	IY	AX	EY	EL	EN	IX
-	75	.	3	9	2	.	1	1	.	1	3	.	.	.	.	.	.	.	.	.	1	.	2	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.
P	25	19	.	19	.	6	.	.	.	.	.	.	.	6	6	.	.	.	.	.	6	.	.	.	.	.	.	.	.	.	.	.	13	.	.	.	.	.
B	12	10	2	.	36	.	.	.	.	.	33	.	.	.	.	.	.	2	.	.	2	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	.	.
T	10	2	2	20	.	8	2	2	.	4	4	2	.	20	4	2	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
D	18	4	4	6	11	.	9	.	.	.	12	.	.	7	2	4	.	16	.	2	2	.	4	.	.	.	.	.	2	.	.	.	2	.	.	.	.	.
K	.	3	.	9	.	21	.	3	.	.	.	3	.	27	21	3	6	3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
G	5	2	.	12	12	7	2	.	.	.	16	.	.	7	14	2	.	14	.	.	2	.	.	.	.	.	.	.	.	.	.	.	5	.	.	.	.	.
F	53	5	.	.	5	.	5	16	.	5	5	.	.	5	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
V	3	.	.	.	3	.	.	.	7	.	20	.	.	.	.	3	7	7	3	17	13	.	13	.	.	.	.	.	.	.	.	3	.	.	.	.	.	.
TH	10	5	.	10	.	.	10	30	.	10	.	5	15	5	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
DH	.	.	14	.	14	.	7	.	.	7	36	.	.	7	.	.	.	7	.	.	.	7	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
S	.	2	.	5	.	3	.	3	.	2	2	63	9	8	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	2	2	.	.	.	.
Z	.	.	.	6	.	.	.	.	.	3	6	62	15	.	6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3
SH	.	.	.	.	.	.	.	.	.	.	.	.	.	33	67	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ZH	.	.	.	.	.	.	.	.	.	.	.	.	.	50	25	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	25	.	.	.	.	.
HH	.	.	.	17	.	.	.	.	.	.	.	.	.	17	.	33	.	.	.	.	.	.	17	.	.	.	.	.	.	.	.	.	17	.	.	.	.	.
M	.	.	4	.	4	.	.	.	.	.	14	.	.	.	.	.	11	36	.	7	4	.	18	.	.	.	.	.	4	.	.	.	.	.	.	.	.	.
N	3	1	2	.	2	.	.	.	.	.	9	.	.	1	.	1	1	66	3	2	1	1	6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
NX	.	5	.	.	5	11	.	.	.	.	5	.	.	.	.	.	.	63	5	5	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
W	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	13	20	.	60	.	.	.	7	.	.	.	.	.	.	.	.	.	.	.	.	.	.
P	.	.	.	.	.	4	.	.	.	.	2	.	.	4	.	.	.	2	.	11	11	.	.	9	2	2	.	42	.	.	2	.	7	.	.	.	.	.
L	2	.	.	4	.	.	4	.	2	.	2	.	.	.	.	.	2	.	.	24	.	13	2	4	4	15	7	4	2	2	.	.	6	.	2	2	.
Y	.	.	.	.	.	.	.	.	.	.	13	.	.	13	.	6	.	.	.	.	.	.	6	.	.	.	.	.	.	.	.	13	50	.	.	.	.	.
UW	.	3	.	.	.	.	.	.	.	.	.	.	3	.	.	.	.	.	.	3	.	3	.	8	13	.	.	.	3	3	3	5	44	10	3	.	.	.
UH	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	8	8	8	8	.	8	15	15	.	.	31	.	.	.	.	.
OW	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	8	15	8	15	8	.	15	8	.	.	.	.	8	.	.	.
AO	.	.	.	.	.	.	.	.	.	.	3	.	.	.	.	.	3	.	.	27	.	7	.	10	.	.	27	3	.	7	7	3	.	3	.	.	.	.	.
AA	.	.	.	.	.	4	.	.	.	.	.	.	.	.	.	.	.	.	.	12	.	.	.	4	.	.	4	15	15	19	.	4	4	15	.	.	.	.	.
ER	.	13	.	8	.	.	.	.	.	.	8	.	.	.	.	.	.	4	.	.	4	.	4	13	4	.	4	.	17	4	.	.	.	8	.	.	8	.	.
AE	.	3	.	3	.	3	3	.	.	.	3	.	.	.	.	.	.	3	.	.	3	.	3	3	6	.	3	.	6	22	17	6	8	.	.	.	.	.	8
EH	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	1	.	1	.	1	4	4	21	15	33	6	2	.	.	.	7
IH	.	.	.	.	.	1	.	.	1	.	1	.	.	1	.	.	.	.	.	.	.	1	10	4	.	1	.	1	.	7	19	16	21	1	.	1	.	12
IY	1	.	.	1	.	2	.	.	1	.	4	.	.	.	.	.	.	.	1	.	1	.	12	.	1	.	.	2	.	.	4	7	51	9	.	.	.	1
AX	.	2	.	2	.	2	1	.	1	.	4	.	.	2	.	.	.	2	.	3	2	.	.	2	5	2	2	5	4	10	5	16	5	23	1	1	.	2
EY	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	20	20	40	.	.	.	20
EL	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	100	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
EN	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
IX	.	.	.	6	.	.	.	.	.	.	6	.	.	6	6	.	.	13	.	.	.	.	.	.	6	.	.	6	.	6	6	.	16	19	.	.	.	6

BF vs. Accuracy and Confusion Matrix
Number of clusters: 75
Distance Metric: EUC
Parametric Representation: ZCC
Data File: TAP (speaker - LC, task - News)

Total Count:	1416				
POS	CL-POS	CL-%COR	PELOST	CL-%COR	BRNCH
1	1.00	20.34	.00	21.05	1.32
2	1.91	31.43	1.00	25.56	1.74
3	2.81	41.95	2.00	31.43	2.17
4	3.70	49.00	3.00	35.03	2.66
5	4.58	53.60	4.00	38.56	3.07
6	5.39	58.26	5.00	41.31	3.52
7	6.20	61.79	6.00	44.07	4.01
8	6.91	64.76	7.00	46.12	4.46
9	7.68	67.94	8.00	48.16	4.90
10	8.52	70.48	9.00	51.20	5.33
11	9.34	73.02	10.00	53.60	5.76
12	10.06	74.93	11.00	55.79	6.20
13	10.74	76.98	12.00	57.91	6.77
14	11.45	78.32	13.00	59.60	7.20
15	12.11	79.66	14.00	61.44	7.67
16	12.80	81.21	15.00	63.06	8.08
17	13.44	82.56	16.00	64.41	8.49
18	14.11	83.83	17.00	65.56	8.88
19	14.75	84.53	18.00	67.66	9.27
20	15.37	85.03	19.00	69.14	9.66
21	16.03	86.02	20.00	70.48	10.02

	-	B	D	G	P	T	K	F	TH	V	DH	S	SH	Z	ZH	HH	DX	M	N	NX	EN	W	R	L	Y	EL	UW	UH	OW	AO	AA	AX	ER	EH	EY	AE	IX	IH	IY
-	43	5	6	.	.	.	.	.	2	1	3	.	.	.	.	.	.	39	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
B	36	.	.	.	5	.	.	.	5	.	.	.	.	.	.	.	.	48	.	.	.	2	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.
D	21	9	4	.	2	2	4	.	7	.	2	.	.	.	.	2	.	35	7	.	.	2	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	2	2	.
G	14	5	.	.	7	2	7	2	12	.	2	.	2	.	2	9	.	33	.	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
P	13	13	.	.	6	8	25	.	6	.	.	.	.	.	.	.	.	25	8	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
T	12	.	2	.	4	12	2	10	16	.	.	.	14	.	.	12	.	8	4	.	.	.	2	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
K	8	3	8	.	3	18	12	9	24	.	.	.	15	.	.	8	.	.	3	.	.	.	.	.	.	.	.	.	.	.	.	.	3	.	.	.	.	.	.	.
F	42	6	.	.	.	.	.	.	11	.	6	.	.	.	.	.	.	37	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TH	25	5	.	.	.	5	.	15	30	.	.	.	.	.	.	.	.	15	5	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
V	7	.	10	.	.	.	.	.	.	7	.	.	.	.	.	.	.	13	23	27	3	3	3	.	.	.	.	.	.	.	.	3	.	.	.	.	.	.	.	.
DH	7	.	14	.	.	.	7	.	7	.	.	.	.	.	.	.	.	67	.	.	.	7	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
S	6	.	3	.	.	8	.	.	6	.	.	31	2	42	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.	.
SH	.	.	.	.	.	.	.	33	.	.	.	.	67	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Z	3	.	.	.	.	3	.	3	3	.	3	32	.	44	.	.	.	6	.	.	.	.	.	.	3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ZH	.	.	.	.	.	.	.	.	.	.	.	.	100	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
HH	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	67	.	.	.	.	.	.	.	.	17	.	.	.	.	.	.	.	.	.	.	.	.	17	.	.
DX	.	.	.	.	.	.	10	.	.	.	.	.	.	.	.	10	10	.	30	.	.	30	.	.	10	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
M	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	43	36	.	18	4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
N	3	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	56	36	.	4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
NX	.	.	.	.	.	.	.	.	.	6	11	.	.	.	.	.	.	42	21	11	5	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	6	.
EN	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
W	.	.	.	.	.	.	.	.	.	7	.	.	.	.	.	.	.	7	20	.	20	40	.	7	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
R	.	.	2	.	4	.	.	.	.	9	.	.	.	.	.	2	.	4	4	.	.	4	20	2	2	2	4	.	.	2	.	7	16	7	.	.	.	7	.
L	.	4	.	.	.	.	.	.	2	11	.	.	.	.	.	.	2	6	2	2	4	6	4	13	.	7	4	7	4	17	.	.	.	.	.	.	.	7	.
Y	.	6	.	.	6	.	.	5	.	.	.	.	.	.	.	6	5	13	.	.	.	.	.	.	.	.	.	.	.	.	.	.	10	.	.	.	13	.	25
EL	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	100	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
UW	.	.	.	.	.	.	.	.	.	.	.	.	.	3	.	.	.	.	.	3	3	5	.	.	3	.	28	.	.	.	.	.	.	.	13	.	13	5	28
UH	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	15	.	15	15	8	.	.	.	15	.	.	.	.	.	8	23	.
OW	.	.	.	.	.	.	.	.	.	8	.	.	.	.	.	.	.	.	.	.	.	.	15	23	.	.	8	.	8	28	.	.	8	.	.	.	8	.
AO	.	.	.	.	.	.	.	.	.	7	3	.	.	.	.	.	.	3	.	3	.	10	10	20	.	10	3	3	13	3	.	3	.	.	.	.	.	7	.
AA	.	.	.	.	4	.	.	.	4	12	.	.	.	.	.	.	.	.	.	.	.	12	8	.	.	4	.	8	.	.	.	.	4	4	4	10	4	15	.
AX	8	1	.	.	1	.	1	.	2	2	.	.	.	.	.	.	.	2	2	1	4	.	10	.	2	4	6	2	2	1	2	13	12	1	2	2	4	20	2
ER	.	.	.	.	.	.	.	.	13	.	.	.	.	.	.	.	.	8	13	.	4	.	4	.	.	4	4	.	.	.	.	8	13	4	4	.	.	17	4
EH	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	7	.	5	.	.	5	1	1	8	5	20	12	2	1	2	20	9
EY	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	20	80
AE	.	.	.	.	3	.	3	.	3	3	.	.	.	.	.	6	3	3	.	.	.	6	.	3	.	.	3	.	.	.	3	14	3	11	3	11	22	.
IX	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	13	.	.	19	.	13	19	6	.	.	.	.	.	.	6	.	.	.	13	6	13
IH	.	.	.	.	1	1	.	.	.	.	.	.	.	.	.	1	.	2	1	.	2	1	1	.	5	.	.	.	.	.	2	7	2	4	.	17	31	19
IY	.	.	.	.	.	.	1	.	1	.	1	.	.	.	1	1	.	4	2	.	2	2	2	1	5	.	2	1	.	.	.	1	8	1	2	.	7	13	80

BF vs. Accuracy and Confusion Matrix
Number of clusters: 75
Distance Metric: COP
Parametric Representation: ZCC
Data File: THP (speaker = CC, task = News)

Total Count: 1416					
POS	CL-POS	CL-%COR	PELOST	CL-%COR	BPNCH
1	1.00	20.7%	.00	22.53	1.34
2	1.92	33.55	1.00	35.03	2.37
3	2.79	42.94	2.00	45.34	3.49
4	3.62	52.33	3.00	52.90	4.56
5	4.37	57.34	4.00	58.97	5.67
6	5.07	62.01	5.00	62.50	6.63
7	5.77	66.17	6.00	65.89	7.51
8	6.40	69.28	7.00	68.57	8.33
9	7.06	71.40	8.00	71.54	9.05
10	7.77	74.29	9.00	72.74	9.75
11	8.43	76.48	10.00	73.73	10.39
12	9.04	78.18	11.00	74.79	11.00
13	9.60	79.52	12.00	75.28	11.54
14	10.20	81.71	13.00	75.99	12.04
15	10.79	82.70	14.00	76.98	12.50
16	11.35	84.68	15.00	77.90	12.97
17	11.87	85.73	16.00	78.46	13.37
18	12.41	86.51	17.00	79.17	13.70
19	12.93	87.43	18.00	79.94	14.06
20	13.47	88.21	19.00	80.44	14.41
21	14.09	88.98	20.00	80.79	14.75

	-	B	D	G	P	T	V	F	TH	V	DH	S	SH	Z	ZH	HH	DX	M	N	NX	EN	W	R	L	Y	EL	UW	UH	DW	AO	AA	AX	ER	EH	EY	AE	IX	IH	IY
-	31	2	35	1	1	.	.	3	1	2	9	.	.	.	.	.	.	12	1	.	1	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
B	17	.	7	.	5	.	5	50	.	.	.	.	.	.	.	.	.	7	2	.	.	.	.	2	.	.	.	.	.	.	.	.	?	.	.	.	.	2	.
D	16	2	18	.	2	4	4	6	4	2	2	.	4	.	.	2	.	28	5	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	4	.	.
G	9	5	5	.	2	5	7	9	9	.	.	.	7	.	5	5	.	26	2	.	.	.	.	2	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.
P	6	6	6	.	.	.	13	.	6	.	13	.	.	.	.	6	19	6	6	.	.	.	.	6	.	.	.	.	.	.	.	.	.	.	.	.	6	.	.
T	8	.	2	.	2	18	2	12	10	.	2	.	8	2	.	20	.	4	.	.	2	2	.	.	4	.	.	.	.	.	.	.	.	.	.	.	.	2	.
K	3	3	3	.	3	21	12	17	12	.	.	.	15	.	3	6	.	.	.	.	.	.	3	.	.	.	.	.	.	.	.	.	3	.	.	.	.	.	.
F	26	.	5	.	.	.	.	16	5	.	5	.	.	.	.	.	.	32	.	.	.	.	5	.	.	.	.	.	.	.	.	.	5	.	.	.	.	.	.
TH	10	.	.	.	.	25	.	45	15	.	.	.	.	.	.	.	.	5	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
V	3	.	10	.	.	.	.	3	.	.	3	.	.	.	.	.	.	27	20	10	.	17	.	.	3	.	.	.	.	.	.	.	.	3	.	.	.	.	.
DH	7	.	7	.	.	.	.	21	7	.	7	.	.	.	.	.	7	36	.	.	.	7	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
S	6	2	.	.	.	5	.	2	2	.	.	28	5	45	2	.	3	.	.	.	.	.	2	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.
SH	.	.	.	.	.	.	33	.	.	.	.	.	67	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Z	.	.	3	.	.	3	.	3	.	.	3	12	.	68	.	.	.	3	.	.	.	3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3	.	.
ZH	.	.	.	.	.	.	25	.	.	.	.	.	75	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
HH	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	67	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	33	.	.	.	.	.	.
DX	.	.	.	.	.	.	10	.	.	.	.	.	.	.	.	20	30	.	20	.	.	.	.	10	10	.	.	.	.	.	.	.	.	.	.	.	.	.	.
M	.	.	.	.	.	.	.	.	.	4	.	.	.	.	.	.	.	46	46	.	4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
N	3	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	62	27	1	4	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.
NX	.	.	.	.	.	.	.	.	.	5	5	.	.	.	.	.	.	42	.	16	21	5	5	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
EN	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
W	.	.	.	.	.	.	.	.	.	.	7	.	.	.	.	.	.	.	13	7	13	20	.	33	.	.	7	.	.	.	.	.	.	.	.	.	.	.	.
R	.	.	.	.	4	.	.	.	.	2	.	.	.	.	2	.	7	9	.	.	4	20	4	.	2	4	.	.	.	.	.	.	18	2	2	.	.	18	.
L	.	2	4	.	.	.	.	.	2	7	2	.	.	.	.	.	2	.	6	.	2	4	11	22	.	7	2	2	5	13	4	4	.	.	.	.	.	.	.	.
Y	.	.	.	.	.	.	.	.	.	.	.	.	6	.	13	13	.	13	.	.	.	.	.	.	.	.	.	.	.	.	.	.	6	.	.	.	.	13	30
EL	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.100	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
UW	.	.	.	.	.	.	.	.	.	.	.	.	.	3	.	.	.	.	5	.	.	.	.	.	.	41	.	.	.	.	.	.	.	3	.	5	10	26
UH	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	15	.	8	23	.	.	15	.	.	.	.	8	8	23	.
DW	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	.	31	.	8	31	.	8	.	.	.	8	.	.	.	8	.
AO	.	.	.	.	.	.	.	.	.	7	3	.	.	.	.	.	3	.	.	.	.	3	20	17	.	10	.	3	.	27	.	.	3	.	.	3	.	.	.
AA	.	.	.	.	.	.	.	.	4	4	.	.	.	.	.	.	4	.	.	.	.	.	42	.	.	.	4	.	4	.	8	8	.	19	.	4	.	.	.
AX	3	.	1	.	.	1	2	.	.	1	.	.	.	.	.	.	.	2	.	.	2	2	12	2	2	5	5	2	5	2	5	5	13	5	1	2	2	17	2
EP	.	.	.	.	.	.	4	.	.	.	.	.	.	.	.	.	8	13	.	8	.	.	13	.	4	.	4	.	.	.	.	.	29	.	4	.	.	13	.
EH	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	4	.	1	9	5	27	16	6	1	4	17	11
EY	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	40	60
AE	.	.	.	.	.	.	.	.	3	.	3	.	.	.	.	.	8	.	3	.	.	.	.	.	.	.	.	6	.	.	6	3	31	3	6	6	.	8	17
IX	.	.	6	.	.	.	.	.	.	6	.	.	.	.	.	.	.	6	.	.	13	.	6	6	6	.	.	.	.	.	.	.	25	.	13	.	.	6	6
IH	.	.	.	.	.	1	2	.	.	.	.	.	.	.	2	.	.	1	.	1	5	1	1	.	6	.	.	.	.	.	.	.	?	2	?	.	1	27	32
IY	.	.	.	1	.	.	2	.	.	1	.	.	.	.	.	.	1	2	.	.	4	.	2	2	10	.	2	.	1	.	.	1	2	1	10	.	2	9	44

BF vs. Accuracy and Confusion Matrix
Number of clusters: 75
Distance Metric: L1).
Parametric Representation: 2CC
Data File: TMP (speaker - CC, task - News)

Total Counts 1416

POS	CL-POS	CL-%COR	PELOST	CL-%COR	BRNCH
1	1.00	19.84	.00	20.76	1.36
2	1.74	26.77	1.00	35.45	2.60
3	2.50	37.22	2.00	47.67	4.15
4	3.20	47.03	3.00	57.34	5.74
5	3.85	52.54	4.00	63.35	7.32
6	4.48	56.29	5.00	67.94	8.77
7	5.11	59.46	6.00	71.47	10.13
8	5.70	62.43	7.00	74.50	11.30
9	6.34	66.10	8.00	76.98	12.35
10	7.00	68.64	9.00	79.17	13.24
11	7.66	70.97	10.00	80.30	14.01
12	8.24	72.81	11.00	81.36	14.63
13	8.79	75.42	12.00	82.56	15.21
14	9.36	76.91	13.00	82.70	15.68
15	9.93	77.97	14.00	83.05	16.10
16	10.48	79.94	15.00	83.26	16.50
17	11.02	81.43	16.00	83.40	16.84
18	11.58	82.06	17.00	83.47	17.14
19	12.10	82.84	18.00	83.76	17.41
20	12.64	83.97	19.00	83.83	17.64
21	13.20	84.46	20.00	83.87	17.86

	-	B	D	G	P	T	K	F	TH	V	DH	S	SH	Z	ZH	HH	DX	M	N	NX	EN	W	R	L	Y	EL	UW	UH	OW	AO	AA	AX	ER	EH	EY	AE	IX	IH	IY
-	42	2	7	.	.	.	.	1	2	.	7	.	.	.	.	.	.	30	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
B	36	.	.	.	7	2	2	.	2	.	.	.	.	.	.	.	.	21	.	26	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
D	21	2	5	.	2	11	2	.	2	2	2	.	2	.	.	2	.	33	11	4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	.
G	14	2	.	.	5	16	5	5	5	.	.	.	7	2	.	2	.	35	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2
P	13	6	.	.	.	25	13	.	13	.	.	.	.	.	.	.	6	13	6	6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
T	12	.	2	.	.	34	2	12	6	.	4	.	14	.	.	.	.	6	4	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.
K	8	.	.	.	9	33	9	12	.	.	.	.	18	.	.	3	.	9	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	8	.	.	.
F	37	.	.	.	.	.	.	5	11	.	.	.	.	.	.	.	.	47	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TH	20	.	5	.	.	10	.	35	15	.	.	.	.	.	.	.	.	5	5	5	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
V	7	.	3	.	.	.	.	.	.	10	3	.	.	.	.	.	.	53	17	.	.	3	.	.	.	.	.	.	.	.	.	.	.	.	3	.	.	.	.	.
DH	7	.	.	.	.	7	.	7	7	.	.	.	.	.	.	.	.	57	.	14	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
S	6	2	3	.	.	5	2	10	3	.	.	45	3	11	.	.	.	2	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
SH	.	.	.	.	.	33	.	.	.	.	.	.	67	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Z	3	.	.	.	.	.	.	9	3	.	6	41	.	26	.	.	.	6	.	.	.	.	.	.	3	.	.	.	.	.	.	.	.	.	3	.	.	.	.	.
ZH	.	.	.	.	.	.	.	.	.	.	.	.	100	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
HH	.	.	.	.	.	.	.	.	.	.	.	.	33	.	.	33	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
DX	.	.	.	.	10	10	.	.	.	.	.	.	.	.	10	.	30	10	20	.	.	10	.	.	.	.	17	.	.	.	.	.	.	.	.	.	17	.	.	.
M	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	70	7	7	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	7	.	.	.	.
N	3	.	.	.	.	.	.	.	.	.	1	1	.	.	.	.	.	75	12	6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.
NX	.	.	.	.	.	.	.	.	.	.	5	.	.	.	.	.	.	69	.	26	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	5	.	.	.
EN	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
W	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	27	.	13	.	7	.	13	.	.	.	.	.	.	.	.	.	.	.	40	.	.	.	.
R	.	.	.	.	4	.	.	2	.	.	.	.	.	2	.	.	2	18	2	.	.	2	24	.	.	16	.	.	.	.	.	7	2	11	.	2	4	.	.	.
L	.	2	2	.	.	.	.	.	2	2	.	.	.	.	.	.	2	17	2	2	.	4	4	17	.	2	6	2	.	8	.	.	.	6	20	.	2	.	.	.
Y	.	.	6	.	.	.	.	13	.	.	.	.	.	.	.	6	.	13	.	.	.	.	.	.	6	.	6	.	.	.	.	.	.	.	.	.	.	6	44
EL	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	100	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
UW	.	.	.	.	.	.	.	.	3	.	.	.	.	.	.	.	.	5	8	.	.	.	3	.	.	.	54	.	.	.	.	.	.	.	.	.	.	16	13	.
UH	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	8	.	.	38	.	.	.	0	0	0	0	.	0	15	.	.
OW	.	.	.	.	.	.	.	.	.	8	.	.	.	.	.	.	.	.	.	.	.	.	8	31	.	.	23	0	.	0	.	.	.	.	8	.	8	.	.	.
AO	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	17	.	.	.	.	10	13	3	.	3	.	.	7	.	.	.	33	10	.	3	.	.	.
AA	.	.	.	.	.	.	.	.	4	4	.	.	.	.	.	.	4	8	.	.	.	4	15	.	.	.	8	4	.	.	.	4	.	.	19	15	0	4	.	.	.
AX	3	1	.	.	.	1	1	.	2	2	1	.	.	.	.	.	1	5	2	.	.	.	6	1	2	2	13	4	.	3	2	4	5	4	27	3	3	2	2	.
ER	.	.	.	.	.	4	.	4	4	.	4	.	.	.	.	.	.	13	.	4	.	.	13	.	.	.	25	.	.	.	.	.	0	.	0	.	.	4	0	.
EH	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	4	2	.	1	.	.	.	1	.	1	.	6	1	.	1	2	4	9	5	82	4	.	16	9	.
EY	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	20	.	.	.	.	.	.	.	.	.	.	40	40	.
AE	.	.	6	.	.	3	.	3	.	.	.	.	.	.	.	.	0	6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	3	20	17	.	17	3	.
IX	.	.	6	.	.	6	.	.	.	.	.	.	.	.	.	.	.	19	13	.	.	.	6	.	.	.	26	.	.	.	.	.	6	.	.	.	6	.	13	.
IH	.	.	.	.	1	2	.	.	.	.	.	.	.	.	.	.	2	4	1	.	.	1	.	.	11	.	11	.	.	.	.	1	6	1	17	.	1	17	21	.
IY	.	.	.	.	.	1	.	.	1	.	.	1	.	.	.	.	1	6	10	.	.	1	2	.	4	.	10	.	.	.	.	1	1	1	1	.	.	10	40	.

BF vs. Accuracy and Confusion Matrix
Number of clusters: 75
Distance Metric: SIG
Parametric Representation: ZCC
Data File: TRP (speaker - CC, task - Navy)

Total Count: 1416

POS	CL-POS	CL-ACCUR	PELOST	CL-ACCUR	BRNCH
1	1.00	22.53	.00	23.09	1.31
2	1.88	31.92	1.00	30.23	1.90
3	2.74	41.31	2.00	34.60	2.39
4	3.50	48.52	3.00	38.63	2.93
5	4.23	52.97	4.00	43.43	3.45
6	4.92	57.20	5.00	47.10	3.94
7	5.60	61.09	6.00	49.44	4.53
8	6.25	64.19	7.00	53.04	5.17
9	6.94	67.51	8.00	55.44	5.70
10	7.66	70.20	9.00	59.04	6.27
11	8.39	72.25	10.00	61.37	6.79
12	9.07	73.16	11.00	62.92	7.28
13	9.60	74.58	12.00	64.27	7.73
14	10.24	76.48	13.00	65.82	8.18
15	10.84	77.75	14.00	67.66	8.63
16	11.45	79.31	15.00	68.57	9.07
17	12.03	80.72	16.00	69.70	9.52
18	12.61	81.50	17.00	70.55	9.91
19	13.18	82.20	18.00	71.40	10.29
20	13.73	82.98	19.00	72.32	10.64
21	14.33	83.55	20.00	73.31	10.99

	-	B	D	G	P	T	K	F	TH	V	DH	S	SH	Z	ZH	HH	DI	M	N	NI	EN	W	R	L	Y	EL	UW	UH	OW	AO	AA	AX	ER	EH	EY	AE	IX	IH	IY
-	42	3	6	.	.	.	.	.	2	1	12	.	.	.	.	.	.	32	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
B	24	.	.	.	10	.	2	.	2	.	55	.	.	.	.	.	.	5	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
D	21	7	4	.	4	7	4	.	2	.	7	.	2	.	.	2	.	20	9	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
G	12	2	.	.	5	7	14	5	2	.	14	.	5	2	.	7	.	23	.	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	2	2	.
P	13	13	.	.	6	6	19	6	6	6	6	.	.	.	.	.	.	6	6	.	6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
T	12	4	.	.	2	22	4	16	8	.	2	.	14	.	.	4	.	6	.	.	.	2	2	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
K	3	.	.	.	9	24	18	12	9	.	.	.	15	.	.	3	.	3	3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
F	26	.	5	.	5	.	.	5	5	.	16	.	.	.	.	.	.	32	5	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TH	25	5	5	.	5	10	.	20	15	.	15	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
V	3	.	3	.	.	.	.	.	.	10	10	.	.	3	.	.	.	17	30	10	.	7	3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
DH	7	.	7	.	.	.	7	.	.	.	29	.	.	.	.	.	.	43	.	.	.	7	.	.	.	.	.	.	.	.	.	.	.	.	3	.	.	.	.	.
S	6	2	.	2	.	8	2	5	3	.	.	43	3	20	2	.	.	.	2	.	.	.	2	.	2	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.
SH	.	.	.	.	.	.	.	.	.	.	.	.	100	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Z	3	.	3	.	.	.	.	3	.	.	3	38	.	41	.	.	.	6	.	.	.	.	.	.	3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ZH	.	.	.	.	.	.	.	.	.	.	.	.	100	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
HH	.	.	.	.	.	.	.	.	.	.	.	.	17	.	.	50	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	33	.	.
DI	.	10	.	.	.	.	.	.	.	.	.	.	.	.	20	10	20	10	10	.	.	10	.	.	10	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
M	.	.	.	.	.	.	.	.	.	.	.	14	.	.	.	.	.	29	50	.	.	7	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
N	3	.	2	.	.	.	.	.	.	1	3	.	.	.	.	.	.	47	39	.	.	4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
NI	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	53	26	16	5	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
EN	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
W	.	.	.	.	.	.	.	.	.	7	.	.	.	.	.	.	.	7	20	20	.	37	.	13	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
P	.	.	.	.	4	.	.	2	.	4	.	.	.	.	.	2	2	4	4	2	.	7	27	2	2	.	13	.	.	.	.	.	2	13	.	.	2	2	.	2
L	.	2	2	.	.	.	.	.	.	17	4	.	.	.	.	.	2	6	4	.	2	4	9	19	.	4	6	4	6	8	.	.	2	.	.	2	.	.	.	.
Y	.	6	.	.	.	.	.	6	.	.	.	.	.	.	.	.	13	.	13	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	6	13	44
EL	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	100	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
UW	.	.	.	3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	5	.	.	0	.	49	.	.	.	.	.	.	8	.	.	.	5	13	15
UH	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	.	.	.	.	38	.	.	.	.	15	23	0	.	.	.	0	.	.
OW	.	.	.	.	.	.	.	.	.	8	.	.	.	.	.	.	.	.	.	.	.	0	.	15	.	23	23	.	0	.	.	.	.	.	.	0	.	0	.	.
AO	.	.	.	.	.	.	.	.	.	13	.	.	.	.	.	.	.	.	.	.	.	0	.	15	.	23	23	.	0	.	.	.	.	.	.	0	.	0	.	.
AA	.	.	.	.	.	.	.	.	.	8	4	.	.	.	.	.	.	4	.	4	.	4	15	12	.	12	8	4	.	4	.	8	.	.	15	.	.	.	.	.
AX	3	1	.	.	.	2	1	.	1	3	.	.	.	.	.	.	2	4	.	.	2	1	11	2	2	1	13	2	4	1	2	12	4	2	1	0	10	5	3	.
ER	.	4	.	.	4	.	.	.	4	.	.	.	.	.	.	.	.	8	13	.	.	0	.	.	.	25	.	.	.	.	4	0	.	0	.	.	4	0	.	.
EH	.	.	.	.	.	.	1	.	.	1	.	.	.	.	.	.	.	1	.	.	.	.	1	.	4	.	.	2	.	2	1	5	17	9	2	12	4	17	19	.
EY	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	20	.	.	.	.	20	60	.	.
AE	.	.	3	.	3	3	.	3	.	.	.	.	.	.	.	.	0	6	.	.	.	0	.	3	.	3	.	.	.	.	.	6	3	6	26	3	11	0	.	.
IX	.	.	6	.	6	.	.	6	.	.	.	.	.	.	.	.	.	6	.	.	13	.	6	.	13	6	.	.	.	.	.	6	.	.	.	.	19	.	13	.
IH	.	.	.	.	4	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	1	.	.	15	.	4	.	.	.	.	5	4	1	2	.	12	20	30	.
IY	.	.	.	.	1	.	.	1	.	1	.	.	.	.	.	.	2	5	.	.	1	.	2	1	12	.	1	.	.	.	.	5	.	.	.	.	7	12	46	.

BF vs. Accuracy and Confusion Matrix
Number of clusters: 86
Distance Metric: EUC
Parametric Representation: ZCC
~~Label file: 0FRM1H.CUS~~ Data file FRM (speaker JB, TASK FRM)

Total	Count:		732			
	POS	CL-POS	CL-%COR	RELOST	CL-%COR	BRNCH
	1	1.00	42.00	.00	42.35	1.10
	2	1.00	51.50	1.00	47.81	1.87
	3	2.60	56.97	2.00	53.96	2.53
	4	3.17	61.34	3.00	58.88	3.21
	5	3.73	64.89	4.00	62.57	4.05
	6	4.47	67.62	5.00	65.03	5.22
	7	5.16	69.81	6.00	67.76	6.06
	8	5.82	71.45	7.00	70.49	6.68
	9	6.49	73.63	8.00	73.77	7.49
	10	6.94	75.14	9.00	76.37	8.12
	11	7.50	75.55	10.00	77.60	8.61
	12	7.89	75.82	11.00	78.83	9.29
	13	8.27	76.78	12.00	79.78	9.91
	14	8.64	77.60	13.00	81.83	10.59
	15	9.22	78.69	14.00	83.20	11.21
	16	9.80	79.64	15.00	83.88	11.71
	17	10.14	80.05	16.00	84.15	12.26
	18	10.51	80.46	17.00	84.97	12.84
	19	11.13	81.15	18.00	86.48	13.45
	20	11.54	81.83	19.00	87.43	14.02
	21	12.09	82.10	20.00	88.39	14.61

	-	P	B	T	D	K	G	F	V	DH	S	SH	HH	M	N	NX	H	L	Y	UW	UH	OW	AO	AA	AH	ER	AE	EH	IH	IY	AX	MH	*
-	77	1	.	3	1	.	1	1	5	3	3	.	.	.	3	.	.	1	.	.	.	.	1	.	.	.	.	.	.	.	.	1	.	1
P	80	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	40	.	.	
B	.100	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
T	.	.	27	4	8	8	.	4	.	16	8	4	4	8	.	.	.	4	.	.	.	.	.	.	.	.	.	.	.	.	8	.	.	
D	.	8	.	42	8	.	.	17	8	.	.	.	8	8	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
K	20	.	.	10	10	10	.	20	10	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	10	10	.	
G	50	.	.	.	.	50	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
F	17	3	.	3	.	17	3	17	7	3	10	.	.	.	3	.	3	.	.	.	.	.	.	.	.	.	.	.	.	.	13	.	.	
V	20	.	.	.	.	.	.	.	27	.	.	.	.	.	20	.	.	.	.	.	.	.	.	.	.	.	.	7	.	.	.	.	.	27
DH	43	.	.	.	.	21	.	.	.	.	.	.	.	.	7	7	.	.	.	.	.	.	.	.	.	.	.	.	.	.	7	7	.	7
S	10	.	.	6	.	.	.	.	.	.	84	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
SH	.	.	.	17	.	.	.	.	.	.	.	67	.	.	.	.	.	17	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
HH	.	.	.	.	.	.	.	.	25	.	.	.	13	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	13	.	50	.	
M	15	.	4	.	.	.	7	.	.	.	.	.	.	15	15	7	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	22	.	16
N	10	5	.	.	.	.	.	.	13	.	.	.	.	.	64	3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	5	.	.
NX	17	.	.	.	.	.	.	.	.	.	.	.	.	17	50	17	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
W	13	.	.	.	.	.	.	.	.	.	.	.	20	.	27	.	33	7	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
L	6	.	.	.	.	.	.	.	10	.	.	.	12	.	6	.	.	35	.	.	.	6	6	.	6	6	.	.	.	.	.	.	.	
Y	.	.	.	.	33	.	.	.	.	.	.	17	.	.	.	.	.	.	33	.	.	.	.	.	.	.	.	.	.	.	17	.	.	
UW	.	6	.	.	.	.	.	.	31	.	.	.	.	.	6	.	.	.	.	6	.	.	.	.	.	6	.	.	.	.	.	44	.	
UH	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	100	.	
OW	.	.	.	.	.	.	.	.	.	.	.	.	8	.	8	.	.	.	.	.	.	.	.	33	.	.	8	.	17	.	.	17	.	8
AO	.	.	14	.	.	.	.	.	7	7	.	.	.	.	.	.	.	7	.	.	.	.	.	.	14	.	29	7	.	.	.	7	.	7
AA	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	17	33	33	8	.	.	.	8	.	
AH	.	.	.	.	5	.	.	.	.	.	.	.	.	.	5	.	.	.	.	.	.	.	.	11	.	21	16	.	.	5	.	5	32	.	
ER	.	2	2	.	.	.	2	.	7	.	4	.	.	4	.	.	.	2	.	.	.	.	.	7	.	.	2	47	.	2	2	.	16	.	
AE	.	6	.	.	.	.	.	.	13	.	.	.	6	.	13	.	6	.	.	.	.	.	.	.	.	.	.	13	6	19	6	.	13	.	
EH	.	.	.	.	.	.	.	.	6	.	.	.	.	.	6	.	.	6	.	.	.	.	.	11	.	6	6	3	6	18	6	.	22	.	6
IH	2	.	.	2	3	.	.	.	.	2	3	2	.	.	2	.	.	2	2	.	.	.	2	.	7	3	5	7	26	10	21	.	.	
IY	.	.	.	8	8	.	.	.	.	.	.	.	.	.	15	.	.	.	.	.	.	.	.	.	.	.	.	.	.	8	54	8	.	.
AX	.	6	.	.	2	.	.	.	16	.	.	.	6	2	6	.	.	10	.	.	.	.	.	6	4	.	4	8	.	.	2	.	29	2	.
MH	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

EM vs. Accuracy and Confusion Matrix
Number of clusters: 128
Distance Metric: EUC
Parametric Representation: ZCC
~~Label file: LAP.ZCC.CWI~~ Data file LAP (speaker LE, task News)

Total Counts: 862

POS	CL-POS	CL-\COP	PELOST	CL-\COP	BRNCH
1	1.00	21.90	.00	22.86	1.03
2	1.94	36.71	1.00	28.70	1.56
3	2.81	43.81	2.00	33.08	2.20
4	3.62	51.51	3.00	39.27	2.81
5	4.38	57.40	4.00	42.75	3.34
6	5.07	62.08	5.00	46.37	3.90
7	5.74	65.11	6.00	50.15	4.49
8	6.31	67.67	7.00	54.68	5.09
9	6.90	69.49	8.00	56.50	5.67
10	7.47	70.54	9.00	60.27	6.19
11	7.93	71.30	10.00	62.69	6.75
12	8.43	72.81	11.00	64.95	7.31
13	8.93	74.17	12.00	66.92	7.90
14	9.39	75.68	13.00	67.98	8.43
15	9.84	76.28	14.00	69.64	8.97
16	10.31	76.74	15.00	70.54	9.48
17	10.76	77.19	16.00	72.05	10.04
18	11.22	77.49	17.00	72.86	10.55
19	11.71	78.10	18.00	73.26	10.99
20	12.19	78.55	19.00	73.72	11.46
21	12.64	79.46	20.00	74.92	11.87

	P	T	K	B	D	G	F	TH	V	DH	S	Z	SH	ZH	HH	M	N	NX	EN	W	R	L	EL	Y	UW	UH	OW	AO	AA	ER	AE	EH	IH	IY	IX	AX
P	43	14	14	.	.	.	.	14	.	14	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
T	14	55	9	5	.	.	.	5	.	.	.	5	.	5	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
K	10	27	10	.	.	9	9	9	.	.	.	.	.	.	.	.	.	.	.	.	.	9	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
B	4	4	.	35	9	13	.	.	4	9	.	.	.	.	.	.	.	4	.	13	4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
D	13	13	.	.	19	.	.	6	13	6	.	.	.	.	.	.	.	13	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	6	.	.	13
G	31	19	6	.	6	6	.	.	6	.	.	.	.	.	.	.	6	6	6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	6	.
F	20	.	20	.	.	.	20	40	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TH	12	18	6	.	6	.	.	47	.	.	.	.	.	.	.	.	6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	6	.	.	.	.
V	7	.	.	7	7	.	.	.	50	.	.	.	.	.	.	7	14	.	.	7	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
DH	50	17	.	.	.	.	.	.	.	17	.	.	.	.	.	17	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
S	4	46	18	.	2	.	7	5	.	.	2	9	.	4	2	.	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.	.	.	2	.	.	.
Z	12	20	4	.	8	4	8	4	.	.	.	.	.	4	4	4	.	.	.	.	.	.	.	.	4	4	.	.	.	.	.	.	.	8	4	.	.
SH	.	17	17	.	.	.	.	.	.	.	.	.	17	33	.	.	.	.	.	.	.	.	.	.	17	.	.	.	.	.	.	.	.	.	.	.	.	.
ZH	.	50	.	.	.	.	.	.	.	.	.	.	.	50	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
HH	.	.	.	.	.	.	.	.	.	.	.	.	.	.	67	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	33	.	.	.	.
M	.	.	.	.	13	.	.	.	13	.	.	.	.	.	.	38	25	6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	6	.	.	.
N	.	2	.	.	23	.	.	.	7	.	.	.	.	.	.	2	22	8	.	27	.	.	.	.	.	.	.	.	.	.	.	.	.	3	.	2	6	.
NX	.	.	.	.	.	.	.	.	22	.	.	.	.	.	.	44	22	11	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
EN	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
W	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	20	.	.	40	.	.	.	.	20	20	.	.	.	.	.	.	.	.	.	.	.
R	5	.	5	.	.	.	.	.	11	.	.	.	.	.	.	11	.	.	.	.	11	.	5	.	.	5	.	6	.	11	11	.	6	.	6	11
L	.	8	.	4	4	.	.	.	25	8	.	.	.	.	.	4	.	.	.	.	.	13	.	.	.	4	4	.	8	.	.	4	.	13
EL	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Y	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	67	33	.
UW	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	9	.	.	.	.	9	.	45	.	.	.	.	18	.	.	.	.	.	.	18
UH	.	.	.	.	.	.	.	.	40	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	20	.	.	.	.	.	.	.	.	.	.	40
OW	.	.	.	.	.	.	.	.	.	25	.	.	.	.	.	25	.	.	.	.	.	.	.	.	.	25	.	25	.	.	.	.	.	.	.	.	.	.
AO	.	.	.	.	6	.	.	.	13	.	.	.	.	.	.	6	.	.	.	6	6	13	13	.	.	13	.	6	.	6	.	.	6	.	.	6	.
AA	.	.	.	.	.	.	.	.	9	.	.	.	.	.	.	3	.	.	.	.	6	9	.	3	.	.	.	9	3	6	3	3	6	.	.	43	.
ER	.	.	.	.	.	11	.	.	.	.	.	.	.	.	.	11	11	.	.	.	.	.	11	.	.	11	.	11	.	33	.	.	.	.	.	.	.
AE	.	.	.	.	.	.	.	.	14	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	14	67	.	.	14
EH	.	.	.	.	.	.	3	.	3	.	.	.	.	.	3	.	3	.	.	3	3	6	.	3	3	.	.	.	10	0	10	27	6	.	6
IH	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.	7	2	.	.	.	4	.	9	27	20	22	7
IY	.	.	2	.	.	.	.	.	2	.	.	.	2	.	.	.	.	.	.	.	.	.	.	2	.	.	.	.	2	.	2	10	86	32	.
IX	.	4	.	.	.	.	.	4	.	.	.	.	.	.	.	.	4	.	.	.	.	.	.	.	4	.	.	.	4	.	.	4	22	30	22
AX	.	.	.	.	.	.	.	2	.	.	.	.	.	.	.	2	.	.	.	2	7	4	4	6	11	6	.	2	.	.	6	.	9	7	7	26

BF vs. Accuracy and Confusion Matrix
Number of clusters: 63
Distance Metric: SIG
Parametric Representation: SPG
Data File: TAP (speaker - CC, task - News)

Total Count: 1416

POS	CL-POS	CL-%COR	PELDST	CL-%COR	BRNCH
1	1.00	31.78	.00	32.06	1.06
2	1.85	49.79	1.00	37.43	1.36
3	2.66	60.24	2.00	43.64	1.72
4	3.42	65.75	3.00	49.86	2.13
5	4.12	70.55	4.00	55.37	2.55
6	4.81	73.59	5.00	61.72	3.02
7	5.47	76.48	6.00	66.31	3.53
8	6.07	78.60	7.00	69.56	4.04
9	6.66	80.01	8.00	71.89	4.52
10	7.25	81.36	9.00	75.14	5.06
11	7.80	82.63	10.00	77.54	6.55
12	8.36	84.25	11.00	79.24	6.05
13	8.94	86.09	12.00	80.16	6.53
14	9.48	87.22	13.00	81.50	7.00
15	10.09	87.92	14.00	82.84	7.48
16	10.54	88.42	15.00	83.69	7.94
17	11.05	89.12	16.00	84.53	8.40
18	11.56	89.76	17.00	85.38	8.83
19	12.03	90.40	18.00	86.09	9.28
20	12.52	90.68	19.00	86.65	9.69
21	12.94	90.75	20.00	87.43	10.09

	-	P	B	T	D	K	G	F	FF	S	N	W	R	Y	UW	UH	OW	AO	AA	AH	ER	AE	EH	IH	IY	AX	EY	AW	AY	OY	EL	EN	DX	+	IX	XX
-	70	5	7	.	2	1	5	2	.	.	3	.	.	3	.	.	.	.	.	.	.	.	.	.	.	1	1	.	.	.	.	.	.	.	.	.
P	25	44	.	.	.	.	6	6	.	.	.	.	.	13	.	.	.	.	.	.	.	.	.	.	.	.	6	.	.	.	.	.	.	.	.	.
B	12	14	36	.	10	.	2	12	.	.	12	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.
T	6	16	2	20	4	6	.	8	6	20	.	.	.	4	.	.	.	.	.	.	.	2	2	.	.	.	.	.	.	.	.	.	.	4	.	.	.
D	9	12	14	5	18	.	4	4	2	.	7	.	2	14	.	.	.	.	.	.	.	4	.	2	2	2	.	.	.	.	.	.	.	4	.	.	.
K	.	3	12	.	.	33	.	3	3	33	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	6	.	.	.	.	.	.	.	6	.	.	.
G	5	7	9	5	21	14	.	2	.	12	12	.	5	7	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.
F	14	11	8	.	7	.	6	10	6	1	11	5	.	11	.	.	.	.	.	.	.	.	1	1	.	1	.	.	.	.	.	.	.	7	.	.	.
FF	.	.	.	.	.	.	.	.	33	.	.	.	.	33	.	.	.	.	.	.	17	.	.	.	.	.	.	.	.	.	.	.	.	17	.	.	.
S	.	4	2	2	1	22	.	2	1	60	.	.	.	5	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	1	.	.	.
N	1	.	1	.	29	1	.	3	.	.	50	.	.	1	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	6	.	1	.
W	1	1	1	.	3	.	3	.	.	.	10	41	1	.	.	.	.	4	1	.	.	1	4	1	.	9	1	.	.	.	10	.	.	1	.	3	.
R	.	2	.	.	.	.	.	.	.	2	11	11	13	2	2	.	.	2	2	.	44	.	2	2	.	.	.	.	.	.	.	.	.	2	.	.	.
Y	.	.	.	.	.	.	.	.	.	13	6	.	.	30	.	.	.	.	.	.	.	19	.	13	13	.	.	.	.	.	.	.	.	.	.	.	.
UW	3	.	.	.	.	.	.	.	.	3	3	.	.	10	13	3	.	.	.	.	.	33	.	13	5	3	8	.	.	.	6	.	.	.	.	.	.
UH	.	.	.	.	.	.	.	.	.	.	.	8	.	.	.	8	.	.	15	.	15	.	.	15	.	38	.	.	.	.	.	.	.	.	.	.	.
OW	.	.	.	.	.	.	.	.	.	.	.	8	.	.	.	.	8	23	.	.	.	.	.	15	.	15	.	.	.	.	23	.	.	.	.	.	.
AO	.	.	.	.	.	.	.	3	.	.	.	40	.	.	.	.	.	3	.	.	7	.	10	.	.	7	.	.	.	.	17	.	3	.	3	.
AA	.	.	4	.	.	.	4	4	.	.	.	.	4	.	.	8	.	8	8	.	.	8	.	8	.	38	.	.	.	.	8	.	.	.	.	.	.
AH	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ER	.	4	.	.	.	.	4	13	.	.	8	.	4	.	4	.	.	.	.	.	33	4	.	8	8	.	.	.	.	.	.	.	.	4	.	4	.
AE	.	3	.	.	3	6	.	3	.	.	.	.	6	3	.	.	.	3	.	.	6	44	6	8	3	3	.	.	.	.	.	.	.	6	.	.	.
EH	.	1	.	.	.	.	.	.	.	.	.	.	2	7	.	.	.	1	9	.	1	31	15	25	1	6	.	.	.	.	.	.	.	1	.	.	.
IH	.	.	.	.	1	1	.	.	.	.	1	.	.	25	.	.	.	1	.	.	1	38	6	17	5	1	.	.	.	.	.	.	.	1	.	.	.
IY	.	1	1	.	1	.	2	.	1	.	1	1	.	26	.	.	1	.	.	.	.	15	1	26	16	5	.	.	.	.	.	.	.	1	.	.	.
AX	1	2	1	.	2	1	.	2	.	1	2	1	.	3	1	8	1	9	3	.	2	7	2	16	2	30	.	.	.	2	.	1	.	2	.
EY	.	.	.	.	.	.	.	.	.	.	.	.	.	20	.	.	.	.	.	.	.	80	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
AW	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
AY	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
OY	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
EL	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	100	.	.	.	.	.	.	.	.	.	.	.
EN	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
DX	.	.	.	10	.	.	.	.	.	.	.	.	.	40	.	.	.	.	.	.	.	.	10	.	.	.	.	.	.	.	.	.	.	40	.	.	.
+	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
IX	.	.	.	6	.	.	.	.	.	13	.	.	25	.	.	.	.	.	.	.	6	6	.	.	13	25	.	.	.	.	.	.	.	.	.	6	.
XX	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

BF vs. Accuracy and Confusion Matrix
Number of clusters: 63
Distance Metric: SIG
Parametric Representation: SPG
Data File: THP (speaker - CC, task - News)

Total Counts 1416

POS	CL-POS	CL-%COP	PELOST	CL-%COP	BRNCH
1	1.00	36.58	.00	36.94	1.04
2	1.75	57.20	1.00	42.00	1.29
3	2.37	66.88	2.00	49.01	1.57
4	2.94	72.95	3.00	55.16	1.90
5	3.44	77.75	4.00	61.30	2.21
6	3.89	80.93	5.00	66.88	2.55
7	4.33	82.63	6.00	71.33	2.92
8	4.72	84.25	7.00	75.00	3.26
9	5.07	85.24	8.00	76.98	3.56
10	5.43	86.23	9.00	79.00	3.89
11	5.75	86.86	10.00	81.50	4.20
12	6.07	87.70	11.00	82.77	4.52
13	6.41	88.56	12.00	83.55	4.81
14	6.73	89.41	13.00	85.03	5.10
15	7.06	89.90	14.00	86.23	5.38
16	7.38	90.32	15.00	86.51	5.67
17	7.69	90.89	16.00	87.01	5.93
18	7.96	91.17	17.00	87.85	6.20
19	8.23	92.02	18.00	88.42	6.47
20	8.50	92.37	19.00	88.91	6.70
21	8.70	92.51	20.00	89.76	6.93

	SIL	UST	VST	LWF	VLF	SIB	VSB	NWS	GLO	LIQ	UWV	AHV	AAV	EPV	AEV	EHV	IYV	OOO
SIL	20	6	13	1	2	.	.	3	3	.	.	.	.	.	.	1	1	.
UST	2	40	12	2	3	18	3	.	4	.	.	.	.	.	1	4	.	.
VST	8	18	39	1	5	2	1	9	10	2	.	.	.	.	1	2	1	.
LWF	20	18	24	24	.	2	.	4	4	.	.	.	.	.	2	.	.	.
VLF	2	2	32	.	7	.	.	16	30	.	.	.	.	.	.	5	2	.
SIB	.	31	3	3	.	19	9	.	4	.	.	.	.	.	.	.	1	.
VSB	.	21	5	.	3	42	24	.	5	.	.	.	.	.	.	.	.	.
NWS	1	1	35	.	3	.	.	58	1	.	.	.	.	.	.	1	1	.
GLO	.	.	6	.	.	6	.	10	52	3	.	.	.	.	10	.	13	.
LIQ	1	2	5	.	.	1	.	10	10	27	1	.	6	20	1	12	4	.
UWV	3	.	.	.	.	.	3	3	10	5	13	3	.	.	33	10	10	.
AHV	.	.	.	.	.	.	.	.	8	.	.	8	15	15	.	30	15	.
AAV	.	.	4	.	3	.	.	.	10	29	.	3	13	3	3	25	7	.
EPV	.	4	8	.	13	.	.	8	.	4	4	.	.	33	4	.	21	.
AEV	.	8	8	.	3	.	.	.	3	6	.	.	3	6	44	8	11	.
EHV	.	2	3	.	1	.	.	1	5	2	.	5	11	2	18	27	22	.
IYV	.	1	4	1	.	.	.	2	26	.	.	.	1	1	25	8	31	.
OOO	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

BF vs. Accuracy and Confusion Matrix
Number of clusters: 63
Distance Metric: SIG
Parametric Representation: SPG
Data File: TAP (speaker - CC, task - News)

Total Count: 1416 POS	CL-POS	CL-%COR	PELOST	CL-%COR	BRNCH
1	1.00	43.00	.00	43.43	1.02
2	1.64	65.40	1.00	49.86	1.24
3	2.12	73.80	2.00	55.23	1.46
4	2.55	78.60	3.00	61.37	1.73
5	2.91	82.63	4.00	66.88	1.96
6	3.24	86.59	5.00	71.96	2.22
7	3.55	86.94	6.00	76.55	2.50
8	3.84	88.42	7.00	79.52	2.74
9	4.07	89.48	8.00	81.07	2.94
10	4.31	90.04	9.00	83.05	3.15
11	4.51	90.68	10.00	84.75	3.37
12	4.70	91.45	11.00	85.88	3.57
13	4.90	92.02	12.00	86.72	3.76
14	5.08	92.58	13.00	88.28	3.93
15	5.26	93.43	14.00	89.19	4.11
16	5.46	93.93	15.00	89.55	4.28
17	5.62	94.35	16.00	89.76	4.44
18	5.77	94.63	17.00	90.61	4.56
19	5.90	95.34	18.00	91.17	4.75
20	6.02	95.69	19.00	91.67	4.87
21	6.10	95.83	20.00	92.37	5.00

	SIL	PLS	VST	FPC	VFP	NAS	LGS	ULV	ALV	ILV	ODD
SIL	70	6	13	1	2	3	3	.	.	2	.
PLS	7	41	12	25	6	.	4	.	1	4	.
VST	8	10	39	3	7	9	12	.	1	3	.
FPC	8	26	12	42	5	2	4	.	1	1	.
VFP	4	11	20	20	16	9	10	.	.	4	.
NAS	1	1	35	.	3	58	1	.	.	2	.
LGS	1	2	5	2	.	10	41	1	8	31	.
ULV	2	.	.	.	2	2	13	13	29	38	.
ALV	.	3	6	.	3	.	29	2	27	31	.
ILV	.	2	4	.	1	2	15	3	26	46	.
ODD	.	.	.	.	.	.	.	.	.	.	.

BF vs. Accuracy and Confusion Matrix
Number of clusters: 63
Distance Metric: SIG
Parametric Representation: SPG
Data Files: TRP (speaker - CC, task - Neut)

Total Count:	1416				
POS	CL-POS	CL-%COR	PELDST	CI-%COR	RPNCH
1	1.00	60.31	.00	60.66	1.02
2	1.46	76.91	1.00	65.32	1.17
3	1.80	84.18	2.00	69.63	1.33
4	2.10	87.71	3.00	74.79	1.52
5	2.33	90.54	4.00	78.74	1.70
6	2.56	92.58	5.00	82.49	1.87
7	2.76	93.43	6.00	85.10	2.04
8	2.93	94.00	7.00	87.22	2.20
9	3.07	94.70	8.00	88.42	2.33
10	3.22	95.34	9.00	89.90	2.47
11	3.33	95.55	10.00	91.24	2.61
12	3.43	95.76	11.00	92.02	2.74
13	3.53	96.05	12.00	92.30	2.85
14	3.63	96.33	13.00	93.50	2.96
15	3.70	96.61	14.00	94.35	3.08
16	3.78	96.75	15.00	94.84	3.18
17	3.85	96.96	16.00	95.13	3.27
18	3.92	97.18	17.00	95.62	3.37
19	3.99	97.39	18.00	95.76	3.46
20	4.05	97.53	19.00	96.97	3.54
21	4.09	97.53	20.00	96.19	3.61

	STP	FPC	NHS	GLO	VHL	SIL	OOO
STP	53	19	6	7	5	8	3
FPC	30	42	5	10	3	6	4
NHS	31	3	58	1	2	1	6
GLO	5	2	10	41	30	1	2
VHL	4	2	2	17	73	.	1
SIL	19	2	3	3	2	70	.
OOO	10	.	.	40	10	.	40

L2: A Machine Transcription

The following is a transcription of some of the sentences in the TAP data set (speaker CC, task News Retrieval). The transcription is the result of segmenting and then labeling with the routines described in this dissertation. The input parameters were the SPG spectrograms, and the labeling was done with the SIG metric and a set of templates very similar to the ones used in the labeling evaluations. (Some hand correction of the template names was done by adding phonetic modifiers to the template labels.) The first column indicates the time in centi-seconds at which either the hand or machine transcription changes. The second column is the hand transcription. The remaining columns indicate, in order of increasing distance (decreasing rating score), the recognized templates. The best score is 50, this is arbitrarily assigned to silences and flaps, which are detected by the segmenter. The ARPABET uppercase phonetic symbols have been used throughout this work.

TIME	PHON	LABELS				
"TELL ME ALL ABOUT REPUBLICANS"						
0		- 50				
10	-					
13	T					
14		T4 43	TZ 40	HH1 39	P2 39	T1 39
18	EH1+1	EH1ED13 44	AA2 43	AO1 41	EH1N11 41	AE1ED11 41
20	EH					
23	EH1L1					
25	L	AX2 43	OW1 41	AO1 41	LZ 40	AE1ED11 40
28		LZ 42	OW2 41	AA1RF14 40	OW1 39	AE6 38
29	M					
30		N3 42	NX1 41	WZ 41	VZ 41	UW1 39
34	IY	EH1ID13 40	AA1RF14 38	AA2 38	ER1 38	IH1L14 36
36		IH8 42	IH6 40	IYZ 39	EH1N11 38	IH1 38
40	IY	IY1 38				
40	IY	IH8 39	IH3 38	IYZ 37	IH1 36	
50		AA1RF14 40	AX1 40	IH1 40	IX1 39	IH5 39
51	+					
52		IX1 41	VZ 40	AX3 40	WZ 38	UW1 38
54	-	B3 37	MZ 37	D1 36	V1 35	TH1 34
56		- 50				
61	+	L1 42	W1 38	OW1 38	UW1 38	
66	AO					
68		AXZ 42	AO1 40	EH1N11 40	AA1 40	AE1ED11 38
72	AO1L1					
78	L					
83	L1+1	L1 39	W1 38			
86	+					
90	AX					
91		OW1 41	AO1L12 40	AX1 38	AE1ED13 38	L1 38
94		IY1 42	UW1 41	L1 40	AA1RF14 40	AE2 38
98	B					
99		B3 42	D1 42	MZ 42	N4 40	D3 38
101		- 50				
106	AE1+1	P1 41	FZ 41	B1HH2 41	T3 39	P3 39
107		EH1ED13 42	IH5 41	EH1N11 41	AO1 39	IH1 39
108	AE					
111	AE1EDRD1	IH1L14 43	AE1ED11 43	AO1 43	AX2 41	UH1 41
121	D	AA1RF14 40	IX1 40	UW1 38	W1 37	OW2 36
123		- 50				
125	T	FZ 34	P1 33	B1HH2 33	P3 31	TZ 30
126		- 50				
127	P	DH1 42	MZ 42	WZ 40	V1 40	N4 40
129		VZ 41	WZ 41	IX1 40	W1 40	UW1 39
132	IX1PF1					
133		P3 41				
135	-					
136		- 50				
144	P	T3 34	TZ 34	ZH1 32	IH8 30	IH3 29
145		N3 39	VZ 38	PZ 37	DZ 36	D3 36
146	AX					
147		IH1L14 40	IX1 38	OW1 38	IH5 38	AE2 37
153		UW1 41	IX1 41	AX3 40	W1 39	WZ 39
154	B					
155		- 50				
159	L	AX3 40	TZ 36	B1HH2 36		
160		AX3 40	NX1 39	M1 38	N1 38	IX1 38
162	IX1L1					
164		N4 40	MZ 39	V1 39	DZ 39	DH1 38
167	-	- 50				
171	K	SH1 40	T4 40	ZH1 39	TZ 37	TH1 37
175	AX1N1	Y1 38	IH1N12 37	IH1N12 36	HH1 36	IY1 35
178	N	- 50				
183	-					
192	Z	SZ 42	S4 39	S5 39	S3 39	S1 38
202		S3 43	SZ 42	S1 42	S5 39	S4 38
206		SZ 43				
209		TH2 38	TH1 38	F1 38	FZ 37	B1 35
212						
213		- 50				
222						
"GIVE ME ALL THE NEWS ON KISSINGER"						
0		- 50				
10	-					
13	G1H1					
14		ZH1 40	T3 40	TZ 36		
16	IH1+1	IYZ 39	IY1 39	IH3 37	IH6 34	ZH1 34
19		IYZ 33	EH1ID13 32	EH1N11 32	IH1 31	ER1N12 30
20	IH					
24	IH1+1	EH1ED13 36	IH8 36	IH3 35	IH1 35	AA1RF14 34

```
  26     V1+1        M1  30        NX1  37        DH1  37        N3  36        W2  36
  29     +
  31     M
  34     IY1+1
  35                 IY2  37       EH1N11  34    IH1  34       IH6  33       EH1ED13  33
  36     IY
  41                 IY1  33       IH0  31       IY2  29       IH6  20
  45     AO1L1
  47                 L1  41        OW1  37       OW2  36       AA1RF14  36   IH1L14  36
  52                 AX2  42       AO1  41       EH1N11  39    AA1  39       AE1ED11  30
  60     L
  65                 L2  41        OW2  41       AA1RF14  39   AE1ED13  30   IH1  30
  67     DH          D1  42        N2  41        B1  40        V1  40        F2  40
  72     AX          OW2  41       AX1  39       AA1RF14  39   IH1L14  30   AX4  37
  76     N           N3  40        M1  30        NX1  37       DH1  37        W2  36
  84     UW          IH0  30       IY2  37       EH1N11  36    IH3  36       IH1  35
  80                 IH5  40       IH6  35       UW1  35
 104                 Z1  37        Z2  36        T1  35        IY1  34        S5  34
 105     Z
 107                 S1  41        S2  40        S4  40        S5  40        S3  39
 100     S
 113     I
 114                 UH2  40       IX1  39       IH3  30       UW1  30       AX1  37
 116     S
 117                 NX1  30       V2  30        M1  37        V1  37        N3  37
 120     AA1N1
 121                 IX1  41       UW1  40       N3  40        AX1  39       W2  30
 126                 IH3  40       AX1  39       AA1RF14  30   IX1  30       IH5  36
 120     N
 129                 N2  39        NX1  39       N1  30        M2  36        W2  36
 135     -           - 50
 141     K           IY2  32       S3  30        IH3  30       S5  30        IH0  20
 143                 ZH1  30       HH1  37       IY1  34       F2  34        Z1  33
 140     IH          IH5  43       IH1  41       IH3  41       EH1N11  41   IH0  39
 163     S
 154                 S5  35        Z1  34        S2  34        ZH1  31
 161                 S3  40        S1  37        S5  35
 164                 S5  39        T3  30        Z1  37        ZH1  37       S2  36
 166     IX          T3  30        TH1  30       T2  30        IH1N17  37    N4  37
 160                 M2  44        N2  44        N4  42        D1  41        D2  41
 169     N
 173     N           - 50
 176     ZH          ZH1  40       T4  30        IY1  35
 100                 ZH1  37       IY2  32
 103     ER
 104                 ER3  42       Y1  41        IH1N12  40   P2  39        IH1N17  39
 101                 V1  41        DH1  40       N1  39        N2  39        W2  30
 105
 109                 - 50
205

"CAN YOU TELL ME ALL ABOUT BUCKLEY"

   0                 - 50
   2                 P1  20       K2  27        TH1  24
   3                 - 50
  10     -
  13     K           T4  40        S5  36        ZH1  36       T1  36        T2  35
  17     AE1N1       EH1N11  37    EH1ED13  37   IH1  36       IH5  35       IY2  35
  25                 IH0  41       IH3  39       IH1  37       IY2  36
  26     N1P1
  27                 N3  41        W2  36        D3  36        DH1  36
  29                 IY2  30
  30     Y
  33     UW
  30     -           B3  30        D1  36        M2  35        N4  35        TH1  34
  40                 - 50
  42     T           ZH1  30
  44                 S5  30        ZH1  37       IY2  35       IH0  34       Z1  34
  47     EH1L1       IH0  42       IH3  42       IY2  40       IH1  39       IH6  30
  49                 EH1N11  44    AA1  42       AA2  42       EH1ED13  41   AE1ED11  41
  54     L           AO1  43       AE1ED11  43   AA2  41       IH1L14  40   AX2  40
  60     M           UW1  35       N3  34        L1  33        V2  31        W1  30
  65     IY          IH0  37       EH1N11  35    EH1ED13  34   IH5  33       IH1  33
  60                 IY2  36
  76                 IH0  36       IH6  35
  77     AO1L1
  01                 OW1  40       AE1ED11  40   IH1L14  30   AX2  39       AA1RF14  39
  04                 AO1L12  41    OW1  40       IH1L14  39   AE2  36       AX4  30
  00                 L1  43        W1  30
  93                 AO1L12  39    AO1  30       OW1  30       AE1ED13  37   AE2  37
  90     L
  99                 L1  42        AO1  39       AO1L12  30    OW1  37
 106     AX
 114                 UW1  30       L1  30        AA1RF14  30   IX1  37       IH1L14  37
 116     B
 117                 - 50
 121                 T3  40        P1  40        T2  40        ZH1  30       B1H12  37
```

```
122     AW      UH1  44      AO1  42      AE1ED11  42     AX2  41      IH1L14  41
133             AA2  42      AO1  41      AA1RF14  41     EH1ED13  40     IH1L14  40
134     O
135             M2  38       N4  38       P1  35         DH1  35        B3  34
137             -  50
139     -
145             T3  37       ZH1  36      IH3  35        B1H12  34       P2  34
146     B1H1
146     AX
147             AO1  45      AX2  44      AE1ED11  43     IH1L14  42      AA2  42
154     -
155             P1  40       M2  38       N4  37         P3  37         TH1  36
157             -  50
163     R       P3  40       B1H12  40    K2  39         P1  38         T2  37
165             P1  41       F1  41       B1  39         TH1  39        D1  38
166     L
167             IX1  41      AX1  40      AA1RF14  39     IH1L14  33     EH1N11  39
170     IY
171             IH3  44      IH8  43      IH6  40
174             IH3  42      IY1  39      IH1N12  38     IH6  37
177             B3  43       B1  41       D1  40         N2  39         M2  39
179
181             -  50
189

"TELL ME EVERYTHING ABOUT ENGLAND"

  0             -  50
 10     -
 15     T       T3  39       T2  37       S5  35
 17             T3  42       TH1  40      D1  39         B1  38         F1  38
 19     EH1L1   IH8  42      AX1  41      IH3  40        IH1  40        AA1RF14  39
 21             AO1  44      AX2  43      AE1ED11  42     AA2  41       EH1ED13  40
 25     L
 29     M       M3  41       UW1  38      W1  36
 36     IY      IY2  37      EH1N11  35   IH1  33        IH8  33        EH1ED13  33
 42             IY2  36      IH8  37
 45             IY2  41      IH8  40      IY1  39        IH3  39        IH6  37
 49             IH3  40      IH6  37      IY3  37        IH1N12  36     IY1  36
 54             P1  36       K2  35       TH1  35        D1  34        B1H12  33
 55     -
 56             -  50
 57             K2  35       D1  34       TH1  34        P1  33        B1H12  32
 58     EH
 59             IH6  41      IH1  41      EH1N11  41     IY2  40       IH3  39
 66             EH1ED13  43  IH5  40      ER1N12  40     EP1  38       AA2  38
 68     V
 69             M2  44       NX1  42      N3  42         M1  42        V1  42
 72     R
 73             EP1  38      P1  37       EH1ED13  36    R3  36         IX1  35
 75             IY2  35      EH1ED13  35  IH1  35        EH1N11  33     IH8  33
 76     IY
 78             IY2  38      IH8  37      IH6  33
 82             T4  42       HH1  41      F2  40         SH1  38       T1  38
 83     TH
 91             P1  38       TH1  37      D3  36         B3  35
 92     IH1N1   EH1N11  44   IH1  41      IH6  40        AX2  40       IH5  39
106             IH6  40      IY1  40      IH3  39        IY2  39       IH8  38
108     NX      NX1  43      N3  42       M1  42         N1  41        M2  39
117     AX1+N1
118             AX1  43      IX1  41      P2  40         AY5  40       IH3  40
121             P2  42       IX1  40      UH2  40        R1D12  39      V2  39
124     -       -  50
128     AX
129             AX1  42      IX1  41      IH5  40        AA1RF14  39    AE1N14  38
135     B       M2  40       DH1  39      D3  39         W2  38        N1  38
137             -  50
141     AW      IH5  40      EH1D13  40   EH1N11  39     AO1  39       AE2  39
147             UH1  42      AO1  42      IH1L14  42     AX2  41       AE1ED11  41
153             EH1N11  40   AO1  40      AE1ED13  40    IH1L14  39     AA2  39
166     T       DX  50
168     I       T4  40       IH1N12  40   IH3  39        IH6  39        HH1  39
163             IY3  43      IY1  40
164     IH1N1
170             IY2  34      IH3  33      IH8  31
172     NX
174             N3  42       N1  41       NX1  41        M1  40        N2  40
180     G
181             B3  38       D1  37       N4  37         M2  36        D3  35
182     L
183             V1  41       N2  41       NX1  41        M1  41        N1  40
189     AX1N1   T4  38       T2  38       T1  36         IH1N12  36     UH2  35
194     N       N4  41       D1  41       B3  40         M2  40        TH1  38
196             -  50
201     -
209     D1H1    TH1  39      P1  38       K2  34
211             -  50
212
```

```
222

"CAN YOU TELL ME EVERYTHING ABOUT DEMOCRATS"

  0                 - 50
 10       -
 14                 T4 39          ZH1 38          F1 37          T2 36          S2 36
 15       K
 18       AE1←1     IH0 42         IH3 40          IY2 40         IH1 38
 19       AE
 20                 IY2 36         IH0 35          EH1N11 34      AA1 33         AA2 32
 23       AE1N1
 24       N1P1
 25                 DX 50
 27       -         IY2 36
 30       UW
 35       D
 36                 - 50
 37       -
 41       T         IY2 34         IH3 33          S5 32          ZH1 32         IH1 31
 43                 HH1 38         T3 38           T4 37          ZH1 36         Z1 36
 46       ←         EH1ED13 43     EH1N11 41       AA2 40         AE1ED11 39     AX2 39
 47       EH1←1
 50       EH1L1
 52                 OW2 42         AX4 39          OW1 38
 53       L
 57       L1←1
 59       M         M1 41          N3 39           NY1 39         V2 38          W2 37
 64       IY1←1     IH0 40         AA1RF14 38      EH1ED13 37     IH1 36         IY2 35
 65       IY
 66                 IY2 35
 72                 IY2 35         IY1 33          IH3 31         IH6 31
 81                 IY1 40         IH6 39          IY2 39         IH3 39         IH1N12 37
 84       IY1←1
 87       ←         D1 40          B1 40           B3 40          V1 36
 90       EH1←1     P2 39          IH3 37          R1D12 37       K2 36          T3 35
 92                 IH1 41         IH0 40          IH5 40         IH3 39         IH6 38
 94       EH
 96                 ER1N12 43      ER1 42          IH5 41         EH1ED13 40     R1 39
 99       V         N3 43          W2 42           UW1 41         M1 40          V2 40
103       R1←1
104                 ER1N12 35      EH1ED13 35      ER1 34         IH5 32         IH1 31
105       IY
107                 IY2 38         IY1 37          IH3 36         IH0 36         IH6 33
108       IY1←1
110       ←         IH3 37         IY1 35          Y1 34          IY2 34         ZH1 33
111       TH
112                 HH1 39         F2 39           T4 37          F1 36          TH2 35
120       ←         EH1N11 41      EH1ED13 41      IH5 41         IH1 40         IH0 40
121       IH
124       IH1N1     IH6 37         IY2 37          IH1 37         EH1N11 37      IH3 36
131       IH1N←1
132                 IY1 41         IH6 41          IH1N12 38      IH3 38         IY2 36
133       NX
135                 NX1 42         N3 42           N2 40          N1 40          M1 39
142       !         IH1N12 40      NX1 39          EH1N12 39      IH3 39         P2 38
145                 P2 38          V2 37           NY1 37         W2 37          R1D12 37
148       AX
150                 IX1 41         AX1 41          IH5 38         IH1N12 38      AX3 38
152       B
153                 - 50
159       B1H1      ZH1 43
160       AE1←1     EH1ED13 41     AX2 41          IH5 40         EH1N11 40      AO1 40
161       AE
164       UH1←1
168       UH        AE2 42         IH1L14 42       AA1RF14 41     AA2 41         AE1ED11 41
170       ←
172       T         M2 41          NH 41           OH1 38         B3 38          P1 37
172       D
174                 - 50
178       -
181                 ZH1 42         T3 40
182       D1H1
183       EH        IH0 40         IH1 40          EH1N11 40      EH1ED13 38     IY2 38
188       EH1N1
189                 IH3 38         AA1RF14 38      IX1 37         AX1 37         IH0 36
190       ←
190       M
191                 V1 42          OH1 41          M1 41          N2 41          D3 41
196       AX        IX1 40         IH5 40          AX1 38         ER1N12 37      ER1 37
200       -
201                 - 50
206       K
207                 AX1 35         IX1 35          IH3 33         AX3 32         R1D12 31
208       R1D1
209                 R1D12 38       HH1 38          V2 34          P2 33
213       AE1X1     P3 37          AX1 37          IH5 37         EH1N12 37      ER3 36
218                 P2 43          AE1N14 40       IH1N12 40      EH1N12 39      IX1 39
```

```
226                 K2  39        P2  38        B1H12  36      IH3  36       R1D12  35
228                 -   50
229                 K2  39        T3  38        TZ  38        TH1  37       P1  37
230     -           -   50
240     S           TH1 40        K2  39        B1H12  39      P1  39       F1  38
242                 S2  41        S4  41        S1  40
255                 TH1 40        B3  39        F1  36        D1  36       P1  36
256
257                 -   50
266

"GIVE US ANY NEWS ABOUT NEPTUNE"

  0                 -   50
 10     G
 12                 IB3 42        D1  41        N2  41        N4  39       B1  39
 16                 ZH1 38        IY2 36        IY1 36        IH3 33
 17     G1H1
 17     IH
 20                 IY2 37        IH8 37
 23                 IH8 38        AA2 35        EH1ED13 34    IY2 34       AE1ED15 33
 25                 IH8 39        EH1ED13 37    IH1 36        IH1L14 36    IY2 35
 26     V
 27                 N3  39        D2  38        IY1 37        NX1 37       M1  37
 31     AX          AE1ED11 40    AX4 40        AA3 40        IH1L14 40    AE2 39
 40                 W1  38        AE6 35
 41     Z
 42                 S2  40        S5  40        S1  39        S3  38       S4  36
 45     S
 47                 S1  42        S2  39        S3  39        S5  37
 53     -           IH1N12 40     IH3 40        B1H12  39     P2  39       N3  37
 55                 TH1 37        P1  36        B3  36        D1  36       N4  32
 57                 -   50
 58     IH1N1       IH5 41        IH1 39        IH1N12 39     IH3 38       EH1N12  37
 64                 IH8 41        IH1 41        AA2 39        EH1ED13 39   IY2 38
 66     N           DX  50
 68                 IY2 39        IH8 35        AA1 34
 69     IY1N1
 73                 IH8 39        IY2 37        IH3 36        EH1N11 34
 77     N
 78                 NX1 43        N3  43        M1  41        N1  40       N2  40
 84     UW          IH3 36        IY2 36        IH8 35        IH1 31
 87                 UW1 29        L1  29        L3  26        IY1 24
107     Z
108                 S3  42        S1  40        S2  37
113     S
116                 S5  42        Z1  40
118     -           D2  40        B1H12  38     IH3 38       IH1N12  38   Z1  37
120                 V1  39        N1  39        D2  38        D1  37       W2  37
122     -
123                 -   50
125     AX          AX1 41        IH5 40        IX1 40        IH1N12  39   IH3 39
133     B           M2  43        D1  42        B3  42        B1  42       N4  40
135                 -   50
138     AW          EH1N11  41    AE2 40        IH5 40        EH1ED13 40   IH1 40
145                 AA3 43        UH1 42        AE1ED11 42    IH1L14  41   AO1 40
153     D           M2  39        B3  38        D1  37        TH1 35       V1  35
155                 -   50
158     N
160                 N2  44        D1  42        N4  41        M2  41       DH1 39
163     EH
164                 IH6 42        IY2 40        IY1 40        IH1 40       IH1N12  39
170                 AX1 42        IH1 41        IX1 39        AA1DF14 39   IH3 39
172     -           D1  43        N4  42        M2  42        B3  41       B1  41
174                 -   50
180     P           T2  36        P1  34        T3  34        K2  33       B1H12  33
188     -
182                 -   50
184     T           ZH1 38        S2  37        S3  36        S1  36       T3  35
187                 T4  39        IH3 38        HH1 37        ZH1 37       T2  37
190     UW1N1       IH3 38        IH1N12 37     IH1N12 37     ER3 36       P2  36
194                 N2  41        NX1 39        M2  38        V1  37       B1  37
199     N
206                 -   50
215
217                 T2  34        K2  34        P1  32        T3  30       P3  30
218                 -   50
225

"TELL US EVERYTHING ON PENNSYLVANIA"

  0                 -   50
 10     -
 15     T           T3  41        T2  40        TH1 39        K2  37       P1  37
 16                 B1  39        F2  38        D1  38        HH1 37       TH2 36
 21     EH1L1       AA2 42        AE1ED11 41    EH1ED13 41    AO1 41       IH1L14  39
 24                 AO1 40        AX2 39        DW1 38        AA3 38       AE1ED11 38
 27     L
```

```
 30     AX
 32             AX4  40        L2  39        AE1ED13  39    AX1  38        AD1L12  38
 36             AX1  39        IH0  37       IH3  37       W1  37         IH1  37
 37     !
 38             Z1  37         ZH1  37       IY1  35       S5  35         T1  35
 40     S
 41             S3  39         S1  39        SZ  38        S5  36
 45             HH1  41        F2  38
 52     S1+1
 53     EH      IH1  42        EH1N11  42    IH6  40       IH5  40        AE2  40
 59             EH1ED13  43    AA2  41
 61             UW1  41        L1  37        W1  36        N3  36
 62     V
 66     R
 67             EH1ED13  38    IH5  36       IH0  36       IY2  34        IH1  34
 69     IY
 70             IY2  39        IH0  34
 74             IY1  41        IY2  40       IH3  38       IH6  38        IH0  37
 76     TH
 77             D1  39         V1  38        D3  37        B1  37         B3  35
 80             TH1  41        F1  40        T3  40        T4  40         TH2  35
 83             IH6  42        P2  41        IH3  41       IH1N12  41     AX1  40
 85     IH1N1
 87             IH6  43        IH1N12  43    IH1  43       IH3  42        EH1N11  41
 93             IY1  40        IY2  39       IH6  39       IH3  37        IH1N12  37
 97     NX
 98             N4  38         D1  37        M2  36        B1H12  36      D2  36
100             N2  44         NX1  43       M1  42        DH1  41        N4  41
106     AA1N1   UW2  40        IX1  39       UW1  39       EH1N12  37     AE1N14  36
110             AX1  42        IX1  40       UW2  40       AA1BF14  40     IH3  39
122     N
123             N2  43         NX1  42        D1  42        M2  41         D3  41
128             -  50
129     -
133             IH3  41        IH6  41       IY1  40       IH1N12  39     T4  38
134     !
136     EH1N1
138             IH1  42        EH1N11  41    EH1ED13  41   IH3  40        IH5  39
142     N
143             N2  42         NX1  42        D1  41        M2  40         N1  39
147             -  50
148     S       T3  42         P3  41        B1H12  41     T2  40         K2  40
149             S3  40         S2  38        S1  38        S5  38         S4  36
155             F1  39         T3  38        D1  38        TH1  37        D2  37
157     AX1L1   AX3  40        IX1  39       W1  38        UW1  38        L1  38
159     L
163     V
164             V1  43         M2  41        M1  40        DH1  40        N4  40
167             -  50
170     EH1N1   EH1N11  42     IH6  41       EH1ED13  40   AX1  40        IH1  39
174             IH0  42        IH6  42       IH3  40       IY2  40        IH1  39
175     IH1N1
178             IY2  38        IY1  38       IH3  37       IY3  36        IH6  35
185     N1P1
186             NX1  42        N2  42        DH1  42       D3  41         M2  41
190     Y
191             IY3  38
198     AX
199             N2  38         N3  37        N4  37        D1  37         V1  36
205             IX1  40        UW1  39       AX3  38       EH1N12  37     AX6  37
208             B1  44         D1  43        B3  43        D2  41         M2  41
210
211             -  50
220
```

"COULD YOU TELL US EVERYTHING ON GOLF"

```
  0             -  50
 10     -
 14     K
 15             K1  38         AX1  37       IX1  36       B1H12  34      AX6  34
 17             D2  39         P1  38        F1  38        B1H12  38      P3  37
 18     UW
 19             EH1ED13  38    IH1  36       AA1BF14  35    ER1  35
 21             IY2  37        IH0  34       EH1ED13  33    EH1N11  33     IH1  33
 23     D
 24             D1  40         B1  40        N2  37
 27     ZH      IY2  35        ZH1  34       IY1  33       IH0  32         S5  32
 30     UW      IY2  37        IH0  35
 32             IY2  36        IH5  33       IH6  32       IH1  32         EH1ED13  32
 38     D       M2  34         D1  34        N4  33        K2  32         B1H12  32
 40     -       -  50
 43     T       IH0  31        ZH1  30
 44             IY1  35        IH0  33       Z1  33        ZH1  33        HH1  32
 48     EH1L1
 48             IH0  42        EH1N11  41    IH1  38       IY2  37        EH1ED13  37
 51             AA2  44        AE1ED11  43   AX2  43       AD1  43        EH1ED13  42
 57     L       L1  37         L2  35        OW2  35       AX4  35
```

```
 62     $X        AX2  44       AO1  43       AE1ED11  42     UW1  41       IH1L14  41
 65               EH1N11  39    EH1ED13  39   L2  38        IH1  38        AX1  38
 72               W1  38        IH0  37       N3  35        V2  34         Z2  34
 73     !
 74               ZH1  40       IY2  38       IY1  38       IH5  36        IH0  36
 75     S
 77               S3  39        S1  38        S2  36        S5  35
 82               T3  37        ZH1  37       Z1  36        S5  36        T1  32
 83     EH
 84               IH5  42       IH1  41       IH3  41       IH1N12  40    EH1N12  37
 93               ER1N12  44    ER1  44       R3  41        R1  40        IH5  40
 95     V
 96               N3  41        M1  40        NX1  40       P2  39        W2  35
100     P
101               IH5  36       IH3  35       IH1  35       ER1N12  34    IY2  34
103     IY
104               IY2  39       IH3  38       IY1  38       IH0  38        IH5  35
107               IH3  41       IY1  40       IH1N12  38    IH1N12  37    T4  36
108     TH
110               NH1  39       F2  38        TH2  37       F1  34
117     IH1N1
118               EH1N11  41    IH1  40       EH1ED13  40   IH5  40        AX1  39
123               IH5  42       IH1  40       IY2  39       IH1N12  38    IY1  38
133     NX
134               NX1  43       N3  42        N1  41        M1  40        W2  39
142     AA1N1
144               UW1  42       IX1  40       UH2  38       L1  38        L3  37
156               IH3  40       AX1  40       AA1RF14  39   IX1  38        IH5  38
158     N
159               N2  41        D1  40        N2  40        NX1  39       D2  39
165     G
167               -  50
169     G1H1      IH0  36       IY2  35       ZH1  35       IH5  34        IH1  34
170     !
171     AA1L1
172               AO1  41       IH1L14  41    AE1ED11  41   AE2  39        AA3  38
185               OW1  41       AO1L12  41    AO1  38       AX2  38        IH1L14  37
186     L
188               UW1  41       IX1  39       W1  37        W2  37         L1  37
190     L1X1
196     F
197               -  50
199               TH1  38
202               -  50
204               TH1  33
213               -  50
214
224
```

"GIVE ME EVERYTHING ABOUT BEEF"

```
  0               -  50
 10     G
 20     G1H1      ZH1  40       T3  38
 22               ZH1  36       S5  35       IY2  33       Z1  32        T3  32
 23     IH
 24               IY2  36       EH1N11  35    IH1  35       EH1ED13  34   IH5  34
 30     !
 31               IH3  38       P2  37        IX1  36       AX1  35        Y1  35
 33               M2  43        V1  42        N4  42        N2  42         M1  41
 34     M
 39               UW1  39       IX1  39       AX3  38       V2  37         M1  37
 41     IY
 42               IY2  36       IH0  31       EH1N11  31
 47               IH0  31       IY1  31       IY2  27
 52               EH1N11  39    IH1  39       IH5  39       AE2  36       EH1ED13  36
 54     EH
 63               EH1ED13  43   EH1N11  42    AX2  41       AA2  41        AA1  40
 66     V         EP1  37       P3  35        EH1EL13  34    IH1  34        IX1  34
 68               V1  43        D3  42        N3  42        N2  41         NX1  40
 71     R
 72               EP1  38       AA1RF14  38   IX1  38       OW2  36        IH1  35
 74     IY        EH1ED13  34   IY2  33       IH1  32       ER1  32        IH0  31
 76               IY2  38       IY1  36       IH0  36       IH3  36        IH5  34
 81     TH        T3  41        ZH1  40       TH1  40       T2  40         F1  39
 83               TH1  41       P1  40        F1  40        B3  39         T2  38
 85               -  50
 87               TH1  43       P1  43        T2  42        F1  42         T3  40
 89               T3  42        K2  41        T2  40        P2  40         P1  40
 90     IH1N1
 91               EH1N11  42    EH1ED13  41   IH5  40       IH1  40        AX1  39
 97               IH0  44       IH1  43       EH1N11  42    IH3  42        IH5  42
101               IH5  39       IY2  39       IH3  39       IY1  39        IH1N12  37
106     NX
108               M1  41        N3  41        NX1  41       N1  41         N2  40
116     AX1N1     UH2  43       IX1  41       AX1  40       AE1N14  39
123     AX1X1
125               P2  40        P3  39        B1H12  39     D2  38         V2  38
```

```
127             AX1  40      IX1  40      UH2  39      AA1RF14  38    AX4  37
130             IX1  41      AX1  41      AA1RF14  41  AY3  39      N1   38
132     B       M2   40      DH1  40      W2   39      M1   39      N1   38
134             -    50
139     AW      T3   41      ZH1  38      TZ   38      B1H12  38    P1   37
140             AO1  41      EH1ED13  41  AX2  40      IH5  40      EH1N11  39
145             AO1  40      EH1ED13  39  IH5  38      IH1L14  38   AE1ED13  38
153     D       M2   36      N4   35      P1   33      DH1  33      D1   32
155             -    50
158     -
162     -
166     B1H1    ZH1  38      B1H12  35    R1D12  34    T3   34      TZ   34
167     IY
168             IH8  37      IH3  37      IY2  37
170             IY2  32      IH3  29
176             ZH1  33      IY1  32      IH3  29      IY3  29
179             IY3  35      B1   33      FZ   33      D1   32      TH2  32
184             IH3  40      IH1N12  39   IX1  38      EH1N12  36    IH1N17  36
185     !
187             FZ   41      B1   41      T4   38      TH2  37
190     F
198             -    50
199             TH1  39      P1   38
200             -    50
202             TH1  37      N2   36      M2   35
206
207             -    50
208             TH1  35      P1   32
209             -    50
216
```

"CAN YOU GIVE ME EVERYTHING ABOUT GANYMEDE"

```
  0             -    50
 10     -
 17     K       S3   35
 20             ZH1  35      T3   34      S5   32      Z1   39
 22     AE1N1   IY2  38      IH8  37      EH1N11  36    IH1  35      EH1ED13  34
 26     N1P1    IH3  39      PZ   37      IX1  36      N3   36      Y1   36
 28             N3   42      VZ   39      DH1  38      W2   38
 31     Y       IY2  41      IH8  38      IY1  36
 35     UW      IY2  39      IH8  34
 39             IH6  39      Y1   38      IY2  38      IY1  38      IH1N12  36
 40     G
 42             B3   40      D1   40      N2   37      B1   37      N4   37
 46     G1H1    S3   37
 48     IH
 49             IY2  39      IH8  36
 53             IH8  41      IH5  40      EH1N11  40    EH1ED13  39    IH1  39
 57     V
 58             MM1  39      IY1  37      Y1   37      FZ   37      PZ   36
 62     M       UM1  41      IX1  41      ER3  39      VZ   38      N3   38
 66     IY
 69             IY2  36      IH6  33
 89             IH8  35      IH6  33      IY1  30
 84     EH
 88             IH1  42      IH6  40      EH1N11  39    AE1ED13  39    IH8  38
 92             IH1  42      EH1ED13  38  IH8  38      AE1ED13  38
 94             EH1ED13  42  IH1  40      AA2  38
 96     V       IX1  37      EP1  37      IH1  37      AA1RF14  36    AE6  36
 98             VZ   40      N3   40      PZ   39      UM1  38      W2   37
101     R       ER1N12  41   ER1  39      EH1ED13  38
104     IY      EH1ED13  34  IH1  32      IY2  31      ER1  29      AA2  29
106             IY2  39      IH8  37      IH6  34
110     TH      FZ   43      MM1  39      TH2  38
119     IH1N1   EH1N11  42   AE2  38      IH6  38
134     NX
135             N3   43      NX1  42      M1   40      N1   39      W2   38
142     AX      IH3  43      IH8  42      IH6  39      IY2  38
144             EH1N11  42   IH1  40      AX1  38      IH1N12  38    IH6  39
147             N3   35      VZ   35      PZ   34      MM1  34      AE6  34
151             IX1  42      AX1  41      AA1RF14  40  IH5  39      IH1N12  38
154     B
156             M2   42      DH1  42      W4   41      D3   40      D2   39
157             -    50
161     !       IH8  42
162     AW      EH1ED13  43  EH1N11  42   AX2  41      AA1  41      AA2  40
166             AO1  43      AE1ED11  42  IH1L14  40   EH1ED13  40   AX2  39
177     D       M2   38      D3   38      DH1  37      N4   37      P1   36
179             -    50
181     G
184             IY2  34      IH3  31      ZH1  30      IH8  30      EH1N11  30
185     G1H1
187     AE1N1   IY1  37      IY2  33      IH8  33
190             IY2  37      IH8  36      EH1N11  35    EH1ED13  34    IH1  34
194             IH8  41      IH1  40      EH1N11  40    IH6  38      IY2  39
198             EH1N11  42   IH1  41      IH8  41      EH1ED13  41    AA2  41
202     N
```

```
203             N3  43        M1  42        D3  41        N1  41        DH1  41
206     IX1N9   EH1N11  38    IH5  37       IH8  36       EH1ED13  36   IH3  36
210     M
211             M1  41        W2  41        NX1  41       N3  40        V2  40
217             IH3  42       IH8  38       P2  38        IX1  38       IY2  37
218     IY
219             IY2  37       IH8  33       EH1N11  32
223             IY1  36       IY2  36       IY3  35       IH6  32
234     D
235             -  50
239     DHH
240             T3  38        T2  38        K2  35        ZH1  35       B1H12  33
241             F2  40        B1  40        T4  40        HH1  39       D1  39
242
252
```