Acoustic-Phonetic Constraints in Continuous Speech Recognition: A Case Study Using the Digit Vocabulary

by

Francine Robina Chen

B.S.E., University of Michigan
(1978)

S.M., Massachusetts Institute of Technology
(1980)

Submitted to the Department of
Electrical Engineering and Computer Science
in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy

at the

Massachusetts Institute of Technology
June 1985

Signature of Author ..Francine R. Chen.....
Department of Electrical Engineering and Computer Science
May 22, 1985

Certified by .....
Victor W. Zue
Thesis Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Departmental Committee on Graduate Students

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JUN 1985		**00-06-1985 to 00-06-1985**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Acoustic-Phonetic Constraints in Continuous Speech Recognition: A Case Study Using the Digit Vocabulary	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Massachusetts Institute of Technology,77 Massachusetts Avenue,Cambridge,MA,02139-4307**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**159**	

Acoustic-Phonetic Constraints in Continuous Speech Recognition: A Case Study Using the Digit Vocabulary

by

Francine Robina Chen

Submitted to the Department of Electrical Engineering and Computer Science
on May 22, 1985 in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Abstract

Many types of acoustic-phonetic constraints can be applied in speech recognition. Shipman and Zue proposed an isolated word recognition model in which sequential constraints are applied at a broad phonetic level to hypothesize word candidates. Detailed acoustic constraints are then applied on a subsequent phone representation to determine the best word from the remaining word candidates. This thesis examines how their model can be extended to continuous speech. We used the recognition of continuously spoken digits as a case study.

We first conducted a feasibility study in which words and word boundaries were hypothesized from an ideal broad phonetic representation of a digit string. We found that strong sequential constraints exist in continuous digit strings and used these results to extend the Shipman and Zue isolated word recognition model to continuous speech.

The continuous speech model consists of three components: broad phonetic classifier, lexical component, and verifier. These components have been implemented for the digit vocabulary for the purpose of exploring how acoustic-phonetic constraints can be applied to natural speech. The broad phonetic classifier produces a string of broad phonetic labels from a set of parameters describing the speech signal. The lexical component uses knowledge about statistical characteristics of the output produced by the broad phonetic classifier to score each of the word hypothesis. Evaluation of this part of the system suggests that it can prune unlikely word candidates effectively.

Nine acoustic features were defined to characterize phones for verifying each of the word candidates. Evaluation of the verifier on the digit vocabulary demonstrates the power of a phone-based representation and of using a few well-motivated acoustic features for describing phones in an acoustic-phonetic approach. In addition to examining the application of speech constraints, evaluation of each of the components indicates that an acoustic-phonetic approach is potentially speaker-independent.

Thesis Supervisor: Victor W. Zue
Title: Associate Professor of Electrical Engineering and Computer Science

Acknowledgments

Many people have contributed to this thesis. I especially want to express my appreciation and thank:

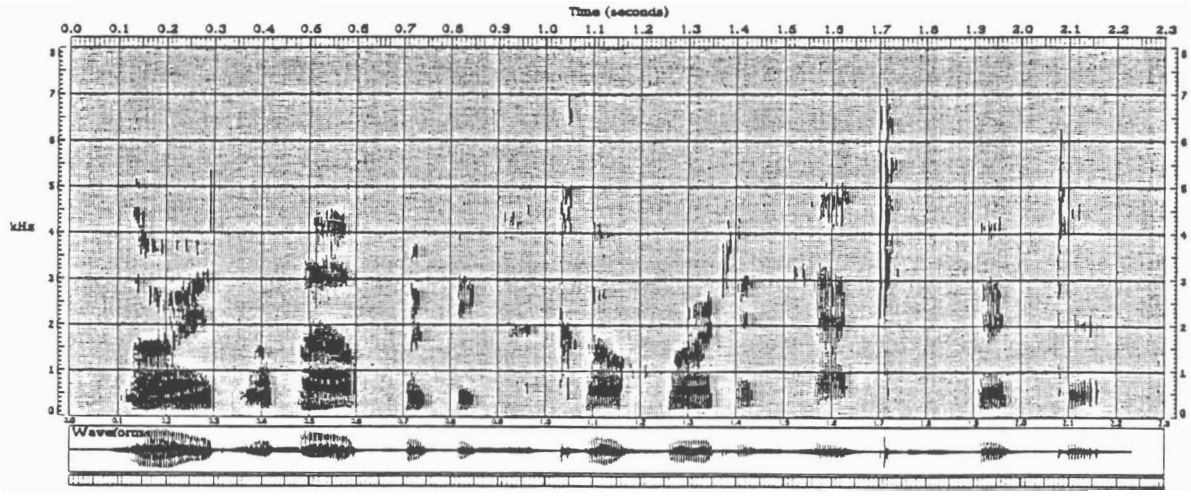
My thesis advisor, Victor Zue, for his guidance, support, friendship, and many Chinese meals. His insight and enthusiasm, and his ability to help me clarify my thoughts and encourage me to try new directions have made working with him a very rewarding experience;

My readers, Jonathon Allen and Kenneth Stevens, for their interest in this work and for their helpful comments on this thesis;

Kenneth Stevens for his advice while I've been a doctoral student and for making the Speech Communication Group a family;

Scott Cyphers, Jim Glass, Lori Lamel, Hong Leung, Mark Randolph, and Stephanie Seneff for discussions which helped to shape this thesis;

Scott Cyphers for helping me develop the rule interpreter, and for taking good care of the bears (our Lisp Machines);



Past and present members of the Speech Communication Group for support and for providing a pleasant and stimulating research environment;

My parents, Robin and June Chen, for their love, support, and understanding;

Leigh, for his love and patience.

Research support for this thesis has been provided by a Vinton-Hayes Fellowship and by contracts from DARPA, monitored through the Office of Naval Research, and by System Development Foundation.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis examines an acoustic-phonetic approach to continuous speech recognition. The approach relies heavily upon low-level speech knowledge—knowledge about phonotactics, the lexicon, allophonic variation, and the duration of different speech units. In this thesis, the constraints provided by each type of low-level speech knowledge were studied and characterized. The constraint information was used to develop components of a speaker-independent, continuous digit recognition system as a research tool. Implementation of the system components allowed a better understanding of how speech constraints could be used in a recognition system.

Speech recognition by computers has possible applications in many areas, ranging from assembly line inspection to airline reservations to aids for the handicapped. Computer recognition of speech could simplify the interaction between humans and computers; one would only need to be able to talk in order to enter information into a computer. We would like speech recognition systems to be speaker-independent, so that a new user is not required to train a system before using it. Furthermore, we would like speech recognition systems to recognize continuous speech, as opposed to isolated words. We use continuous speech, not isolated words, when we speak; therefore, a continuous speech recognition system is more user-friendly. Continuous speech recognition systems have the added advantage that users could enter infor-

mation into a computer more quickly, since the speaking rate is higher in continuous speech.

In the past, researchers have expended much effort developing and refining recognition systems based chiefly on engineering techniques. This is primarily a reflection of relatively primitive and incomplete knowledge about acoustic, phonetic, and other low-level characteristics of speech. However, we now understand these characteristics more fully than we did a decade ago. By exploiting what knowledge we have about speech and then pushing our knowledge further, better and more advanced speech recognition systems may be developed.

1.1 Speech and Speech Knowledge

Speech sounds are produced as air flows through and resonates in the vocal tract. Different speech sounds are due to different configurations of the vocal tract, each of which is associated with a set of resonant frequencies. In addition, different sounds are produced depending on the excitation source. The excitation may be at the glottis and/or at a constriction(s) in the vocal tract. When the excitation is at the glottis, the vocal folds can remain open to produce aspiration, as in /h/, or the vocal folds may vibrate rhythmically to produce voiced, periodic sounds, such as vowels and nasals. The peaks in the spectrum of a voiced sound are called formants and are labeled as F_i for the i^{th} formant. Thus, for example, the formant with the lowest frequency is called the "first formant" and is labeled F_1. When the excitation is at a constriction in the vocal tract, noisy aperiodic sounds (e.g., /s/, /š/) are produced. More than one source may be present during production of a sound. When both a noise and voicing source are present, voiced consonants (e.g., /v/, /z/) are produced. Many speech scientists (e.g., Chomsky and Halle, 1968; Jakobson, Fant, and Halle, 1952) have described speech sounds in terms of these characteristics, that is, voiced or unvoiced characteristics, and other characteristics

such as frontness of a vowel.

Because the rate the vocal tract articulators can move is limited, the articulators are not instantaneously positioned to produce each sound. Consequently, each sound is affected by its neighbors, or context; this is called *coarticulation* (see for example, Heffner, 1950). As an example, there are at least two kinds of /k/; each is illustrated in the spectrograms of Figure 1.1. The particular realization of a /k/ depends on the adjacent vowel: The /k/ on the left is followed by a front vowel and is called a "front /k/," and the /k/ on the right is followed by a back vowel and is called a "back /k/." Note the higher burst frequency of the front /k/; this is due to the constriction formed with the hump of the tongue against the roof of the mouth being positioned farther forward in a front /k/ than in a back /k/.

Due to differences in rate of articulatory movement, some sounds are relatively stable in time compared to others. For example, an /n/ is much more stable than an /r/ (compare the /n/ and /r/ in Figure 1.2). The transition from vowel to nasal, primarily due to the velum lowering to couple the oral and nasal cavities, is rapid because the velum can be quickly lowered. Once the transition is made, the nasal is stable for its duration. In contrast, an /r/, produced by retroflexing the tongue, usually shows movement throughout its duration. Since the tongue cannot move as quickly as the velum, the time it takes to retroflex the tongue to produce an /r/ can be observed in a spectrogram as gradual lowering of F_3.

In many languages, only limited sequences of sounds are allowed (Sigurd, 1970; Shipman and Zue, 1982). For example, in an English syllable beginning with three consonants, the first consonant must be an /s/, the second either /p/, /t/, or /k/, and the third either /l/, /w/, or /r/. Furthermore, not all combinations of these sounds are allowed. Thus given a sequence of sounds, one can deduce whether or not it could be a word in a specified language.

The examples in this section have briefly introduced some low-level speech characteristics. These characteristics can be organized as low-level speech knowl-

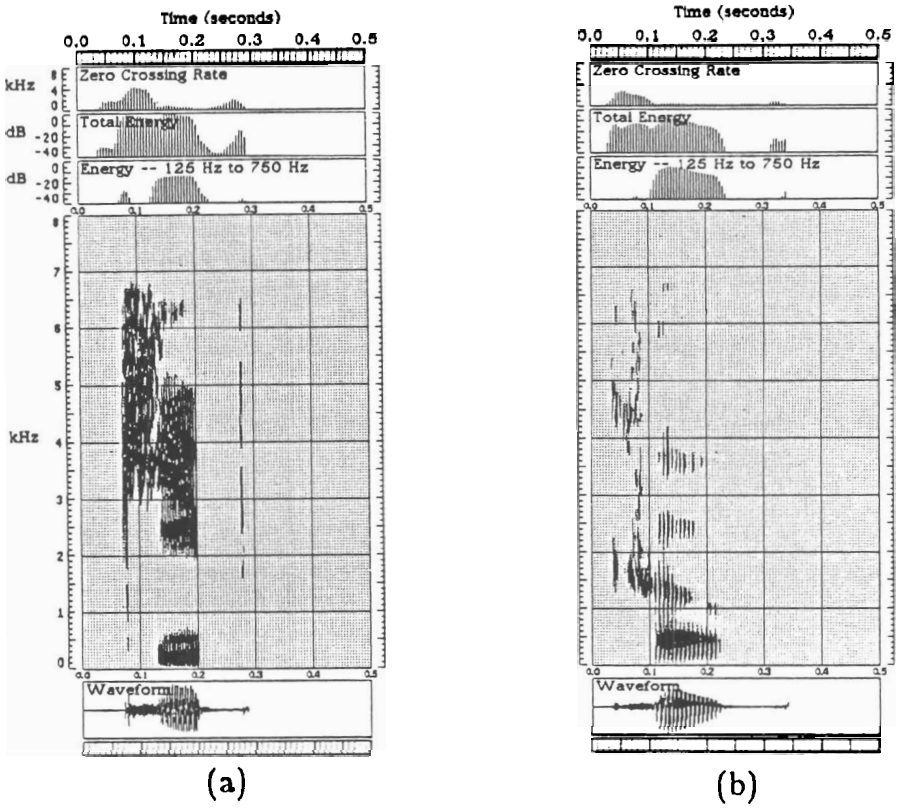


Figure 1.1: Front and back /k/'s: (a) front /k/ in "keep" (b) back /k/ in "coop"

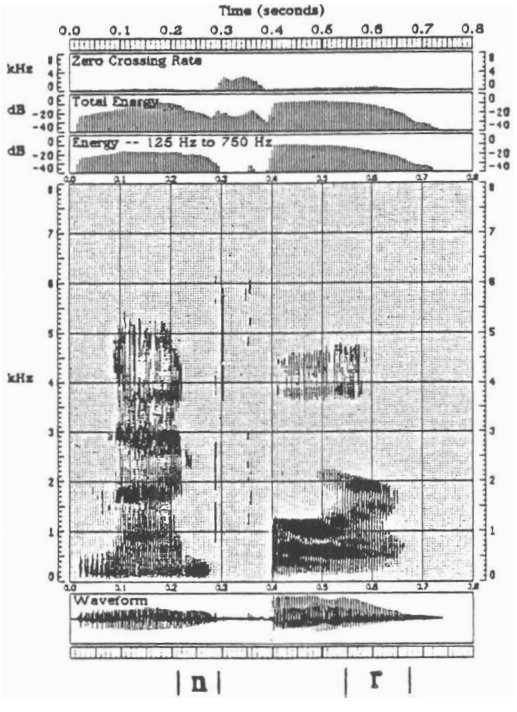


Figure 1.2: Rate of articulatory movement for /n/ and /r/ in the digit string "nine four"

edge: knowledge about acoustic characteristics of sounds, coarticulation, duration of sounds, and phonotactics. The formulation of this knowledge into speech constraints and the use of these constraints in recognition of natural speech were explored in this thesis.

1.2 Speech Recognition Systems

The speech recognition systems developed during the past 15 years have used varying amounts and types of speech knowledge. Some systems, such as those based on template-matching, use little, if any, speech knowledge. In contrast, the explicit use of speech knowledge forms the basis of the FEATURE system developed at Carnagie Mellon University. In this section, some benchmark recognition systems are described. (Many excellent reviews of major recognition systems have been written, such as by Lea, 1980 and Klatt, 1977.) Systems which use little explicit low-level speech knowledge are described first, followed by descriptions of systems which progressively use more speech knowledge in an explicit manner. At the end of this section, the use of speech knowledge by each of the described systems is compared.

1.2.1 Template Matching

Most speech recognition systems presently on the market are based on a mathematical approach which combines template matching with a time-alignment procedure known as dynamic time warping. These systems perform with over 95% accuracy on speaker-trained isolated-word and limited-vocabulary connected-word tasks (Kaplan, 1980; Doddington and Schalk, 1981). In template matching, each recognition unit, for example, a word, is represented by at least one template, created from a set of training utterances (Rabiner, 1978). Each template is composed of a sequence of patterns in time and each pattern, in turn, is some parametric

representation of speech. For example, a template could be composed of the sequence of linear prediction coefficients sampled every 5 msec throughout a training utterance.

Dynamic time warping is a method based on dynamic programming for nonlinearly aligning two sequences. Some systems constrain the amount of time compression and expansion allowed; for example, Myers and Rabiner (1981a) limited time normalization to a ratio of 2:1 between the reference and test templates.

To recognize a word, an isolated word recognizer compares the input signal against each of the stored word templates using dynamic time warping and a predefined distance metric (e.g., Itakura's distance metric, see Itakura, 1975). The best alignment between the input signal and each template is found. The input word is classified as the word corresponding to the best matching template, that is, the word with the smallest distance.

Many constraints on isolated words are not present in continuous speech, and, as a result, extension of isolated word techniques to continuous speech is not straightforward. In isolated word recognition, the word endpoints are "known." In contrast, in continuous speech recognition, the word endpoints are unknown and coarticulation occurs between words. Because handling unknown endpoints requires extra computation, systems developed to extend template matching to continuous speech (e.g., Sakoe, 1979; Kato, 1980; Myers and Rabiner, 1981a; Myers and Rabiner, 1981b) have used very small vocabularies. Defining templates which differentiate among similar words is not an easy task using conventional template representations, such as linear prediction coefficients, because such representations do not adequately emphasize fine differences. Defining templates for continuous speech is an even more difficult task because one must have a framework to describe coarticulation across word boundaries. Conventional template representations do not lend themselves to descriptions of coarticulation because they represent an acoustic event and not an acoustic manifestation of a phonetic event. For reasons such as

these, extending template matching from isolated word recognition to less restricted speech recognition tasks has proven to be difficult.

1.2.2 Harpy

The HARPY system (Lowerre, 1977; Lowerre, 1980) demonstrated the best performance of all the continuous speech recognition systems developed under the ARPA project[1], with less than 5% semantic error on sentences from a highly constrained grammar. Quasi-stationary segments derived from simple parameters, called "zapdash" parameters, were "mapped to the network states based on the probability of match...by use of phonetic templates" (Lowerre 1977). Along with juncture rules, 98 templates were used to represent all possible allophones and speaker variations. A score was assigned to each node in the network based upon how well the zapdash parameter values in a segment of speech matched the templates. Finally, the finite state graph was searched using a heuristic but efficient search algorithm, called a beam search, to find the best path (subject to constraints used in the algorithm) through the network and the corresponding best sentence.

The developmental effort in Harpy concentrated on using high-level knowledge. As a result, the Harpy system relies heavily on higher level language constraints, such as its constrained grammar. Innovations introduced in the system include higher level processing of the input signal, a precompiled network embodying many forms of higher level knowledge, and the beam search algorithm. In contrast to its well-developed high level processing, Harpy's relatively primitive front end performed only rough segmentation based upon zero crossing rates and peaks in smoothed and differenced waveform parameters (the "zapdash" parameters).

The success of the Harpy system demonstrates that higher level constraints are

[1]During the early seventies, the Advanced Research Project Agency (ARPA) sponsored a five year project, involving approximately ten research groups, to study the feasibility of building systems for speaker-independent, continuous speech understanding.

useful in speech recognition. In addition, it illustrated how constraining a task can be used to reduce a problem to manageable proportions. However, using only higher level constraints can reduce the task too much to be of general use. In fact, a general criticism of Harpy is that its grammar is so constrained that it is not habitable. Even so, the Harpy system exemplified how higher level language constraints can be applied to the recognition problem.

1.2.3 IBM

IBM has developed two benchmark systems: a speaker-trained continuous speech recognition system and a speaker-trained isolated word recognition system. The IBM continuous speech recognition system (Jelinek, 1975; Jelinek, 1976; Jelinek, 1981) has a recognition accuracy of 91% correct (Jelinek, 1980) on words contained in sentences in the 1000 word vocabulary of the *Laser Patent Text*. The isolated word recognition system has a recognition accuracy of 95% on a 5000 word office correspondence vocabulary (Bahl et al., 1983).

Both systems are based upon Hidden Markov Modeling, a statistical technique which IBM has applied to model specified speech units. In both systems, each word in the lexicon is represented as a sequence of phonemes in a finite state graph. The phonemes, in turn, are represented as a sequence of templates which attempt to capture variations of all phonemes (allophones) in the language. In the continuous speech recognition system, the templates are computed from DFT (Discrete Fourier Transform) coefficients, which evenly weight the spectrum even though the information content in the spectrum is not uniform. In contrast, in the isolated word recognition system, principal component values derived from the DFT representation are used as input from the front end, weighting the DFT coefficients in accordance with characteristics of speech sounds.

Hidden Markov Modeling characterizes the speech signal by statistical methods; thus, the system can be trained without human input. However, Hidden Markov

Model systems require large amounts of speech data and computation for training. In particular, training a continuous speech recognition system to one speaker may require many hours of speech and many hours of computer time, making it unacceptable for general use. In contrast, the isolated word recognition system requires much less training than the continuous recognition system, although it still does require hours of computer time on IBM's largest computer. By constraining the task to isolated word recognition, IBM made the training and computational costs for a Hidden Markov Model recognition system manageable.

1.2.4 Hearsay II

The research effort of CMU's Hearsay II system (Erman and Lesser, 1980; Lesser et al., 1975), developed under the ARPA project to recognize continuous speech from several speakers, was directed at developing and studying the interactions of knowledge sources. Each knowledge source contained "knowledge," such as speech descriptions, needed by the recognition system to solve a particular recognition task. The knowledge sources were modular, which allowed easy modification of knowledge, but the cost was slower recognition and a more complex control structure. In the resulting system configuration, independent, parallel, knowledge sources communicated through a multilayer global blackboard. The "layers" of the blackboard included: segment labels, syllables, proposed lexical items, accepted words, and partial phrase theories. A knowledge source was activated when new information on the blackboard caused its preconditions to be met. An activated knowledge source would attempt to provide information to a higher level (bottom-up analysis) or lower level (top-down analysis).

Much less emphasis was placed on development of low-level knowledge sources concerned with acoustic-phonetics relative to higher level knowledge sources concerned with syntax and semantics. For example, the Hearsay II segmentation component simply used template matching and a well-developed algorithm (Itakura's

distance metric) to assign to each segment a label corresponding to one of 98 possible templates. The performance of the low-level knowledge sources was poor: the segmentation component assigned the correct label as its first choice only 42% of the time, and the word hypothesizer hypothesized the correct word to be within the top 50 candidates out of a possible 1000 words only 70% of the time (Klatt, 1977). But higher level syntactic and semantic knowledge sources allowed recovery from these errors by top-down word hypothesization. Thus, Hearsay II demonstrated the utility of knowledge to drive a recognition system, especially the use of higher level knowledge sources.

1.2.5 HWIM

The HWIM system (Hear What I Mean), developed under the ARPA project at Bolt, Beranek, and Newman, had fixed components in a roughly hierarchical structure. The system initially segmented and labeled the speech signal as a set of phonetic transcription alternatives arranged in a "segment lattice." The purpose of the lattice was to avoid fatal segmentation errors by allowing ambiguity. The first choice accuracy of assigning the correct phonetic label (out of 71 possible labels) was only 56% (Schwartz and Zue, 1976).

The designers of HWIM introduced several interesting ideas for speech recognition. These included word verification at the parametric level, phonological rules for building a lexical decoding network, and top-down verification using the context and position of a word (Cook and Schwartz, 1977). HWIM's parametric word verifier scored word candidates at the parametric level by matching frames of a word candidate with frames of the corresponding synthesized word. Speech knowledge was used in developing the front end descriptors for each of the phonetic labels and in developing phonological rules. Language knowledge was used to find the best path through the word lattice and in top-down verification. However, the methods used to incorporate speech knowledge into the system were not fully explored. For

example, "intuitive human guesses, not statistically measured estimates" of likelihoods for word pronunciations were used to evaluate the system due to lack of time (Wolf and Woods, 1980). This was done because obtaining sufficient statistics to produce meaningful pronunciation likelihoods is difficult.

1.2.6 FEATURE

The FEATURE system (Cole et al., 1982), which recognized the letters of the alphabet, used acoustic features of speech for discriminating among speech sounds. To recognize a word, the system extracted acoustic features from parameters between four temporal anchor points. These anchor points were chosen to take advantage of the monosyllabic structure of most of the lexical items. The parameters used in the system, motivated by acoustic-phonetic knowledge of speech, include formant frequencies, fundamental frequency of voicing, zero crossing rate, total energy, low frequency energy, mid-frequency energy, and high frequency energy.

FEATURE's average recognition rate was 89.5% correct when tested on 10 male and 10 female speakers. In light of the difficult vocabulary (many of the letters of the alphabet sound similar), FEATURE's performance demonstrates that acoustic-phonetic knowledge can be useful in speech recognition. However, extensibility of the FEATURE algorithm to continuous speech is uncertain. The recognition algorithm took advantage of the monosyllabic nature of the vocabulary words and the analysis was done from four anchor points. In addition, many hours of a speech scientist's time were required to develop features for identification in the limited vocabulary; many more hours would be needed to develop features for all contexts.

1.2.7 Speech Knowledge in Recognition Systems

We have seen several approaches to speech recognition, each using varying amounts of speech knowledge. Template matching techniques use constraints to define a manageable task (e.g., speaker-trained and isolated word tasks) and are

relatively easy to develop because they are mathematically well defined. In template matching systems, only a very small amount of speech knowledge is used relative to the potential amount of knowledge that could be used. In these systems, speech knowledge is usually incorporated by limiting the amount of time compression/expansion, reflecting some knowledge about limits on speech rate variation. Speech knowledge has also been incorporated through choice of recognition unit; for example, Rosenburg et al. (1983) developed a system which uses the demisyllable as the recognition unit.

The Harpy system used more speech knowledge than template matching systems. Some speech knowledge was needed to develop the "zapdash" parameters, and template selection required some speech knowledge. However, most of the knowledge used was higher level language knowledge, such as grammar and syntax.

The IBM systems showed how some speech knowledge combined with well defined mathematical techniques can successfully be used for recognition. Speech knowledge was explicitly used in defining the sequence of templates representing a word and in defining allowable word pronunciations due to coarticulation. Statistical characterization of the speech signal provided much strength to the systems. However, this characterization was performed without explicit use of knowledge about speech and also required a lot of training data.

The importance of language constraints in speech recognition and how such constraints can be used to recover from low-level errors was demonstrated by Hearsay II. HWIM and FEATURE illustrated that low-level speech constraints may be important in speech recognition. In particular, HWIM exemplified how constraints on phonological variations may be important, and FEATURE exemplified how acoustic-phonetic constraints may be important.

Each of these systems has contributed to our understanding of how to use speech knowledge in speech recognition. However, we still need to understand better the constraints provided by different types of speech knowledge, especially low-level

speech knowledge, and how to cohesively use the constraints in the speech signal for recognition.

1.3 Problem Statement and Overview

This thesis studies the application of low-level constraints in the speech signal to continuous speech recognition, particularly the task of recognizing continuous digits. Lexical, durational, acoustic, phonetic and allophonic speech constraints are examined and the utility of these constraints is tested on natural speech. In contrast, high level knowledge about the language, such as grammar, syntax, and semantics, is not addressed. The investigation was divided into three parts:

1. develop a continuous speech recognition model based upon constraints in the speech signal

2. implement components of the model, making the required modifications to accommodate variabilities in the speech signal

3. explore the use of detailed acoustic analysis of phones for verification of word hypotheses

Chapter 2 develops a continuous speech recognition model which relies heavily upon speech knowledge and is based on sequential constraints, as used by Shipman and Zue (1982). Shipman and Zue's work in sequential constraints for isolated word recognition is described first. A feasibility study which was conducted to show that strong sequential constraints exist at the broad phonetic level in digit strings is described next. The results of the study indicate that for a continuous speech recognition task such as the digits, a recognition system can initially process continuous speech at the more robust broad class level, rather than at the detailed level of the benchmark systems. This philosophy was used in the development of the recognition model.

Other low-level speech constraints which may be useful in a recognition system are described next. The use of additional constraints provided by the low-level properties of the speech signal performs two functions. First, the additional constraints "counteract" the loss of word endpoint constraint in continuous speech. Second, the additional constraint may allow "higher" level language constraints to be relaxed. For example, the grammatical constraints in Harpy may be relaxed to form a more habitable grammar.

In the last section of Chapter 2, the Shipman and Zue model is extended to continuous speech. The general organization of the model is presented, and the incorporation of constraints in the speech signal into the model is discussed.

The next two chapters describe the implementation of the components in the recognition model. Chapter 3 considers how speech constraints can be applied in the broad phonetic classifier and lexical component. Chapter 4 explores an acoustic-phonetic approach to verification of word hypotheses in a word lattice.

In Chapter 3, refinements to the continuous speech recognition model to handle the variations which occur in natural speech due to interspeaker and intraspeaker variations are described. These variations can lead to recognition errors unless the recognition algorithm is developed to explicitly deal with them. In addition, contextual and coarticulatory variations occur in natural speech and must be incorporated into the algorithm. For example, a person may pronounce "five" with or without a /v/, as in [fa$^{\nu}$ve$^{\nu}$t$^{\neg}$] ("five eight") and [fa$^{\nu}$na$^{\nu}$n] ("five nine"), respectively. A method for segmentation and labeling by characterizing speech in terms of acoustic features by the broad phonetic classifier is described. Then an algorithm for applying sequential constraints to natural speech using knowledge of front end characteristics is developed. Finally, an investigation of the application of path, allophonic, and durational constraints is presented.

Chapter 4 describes how the knowledge gained from low-level constraints is used when performing detailed acoustic analysis. In particular, verification of word

hypotheses using a phone-based representation was explored. A set of acoustic features for characterizing and identifying the phones in the digit vocabulary are described. A method for using information from the features to score each phone hypothesis is presented and then the method is evaluated.

Each of the implemented components was evaluated. The broad phonetic classifier was evaluated by comparing its output to a hand-labeled phonetic transcription. The lexical component was evaluated using output from the broad phonetic classifier, since the main contribution of the lexical component is in using the characteristics of the broad phonetic classifier in evaluating word hypotheses. The verification component was evaluated through incremental simulation, so that errors due to the verifier could be separated from errors due to other components.

Chapter 5 justifies the assumptions made in the thesis and outlines the contributions of the thesis. This chapter discusses the most prominent assumption, that an acoustic-phonetic approach has many advantages over other approaches. Additionally, the characteristics of the digit vocabulary and how the choice of vocabulary affects the study are described. The utility of a preprocessor, especially an acoustic-phonetic preprocessor, the use of segments and sequential constraints in recognition, and computational considerations are also discussed. Finally, the contributions of the thesis are outlined and suggestions are made regarding ways the research can be refined and extended.

Chapter 2

A Speech-Knowledge Based Recognition Model

This chapter discusses a continuous speech recognition model which uses speech knowledge to constrain the recognition task during each processing step. The philosophy of the model is general enough to serve as a basis for developing large vocabulary continuous speech recognition systems using low-level speech knowledge. Since the model was implemented on a limited task, the digits, some of the background work is based on analysis of only the limited vocabulary. Suggestions on how the model may be relevant to larger vocabularies are discussed in Chapter 5.

The model follows the work of Shipman and Zue, described in Section 2.1, on sequential constraints in isolated words. The feasibility of a sequential constraint based approach for continuous speech was analyzed with a modeling experiment, described in Section 2.2. This study showed that strong sequential constraints on a broad class representation of continuous speech can be used to specify word hypotheses and corresponding word boundaries for the limited case of digit strings. However, sequential constraints in continuous speech do not provide as much constraint in identification of the utterance as sequential constraints in isolated words, because definite word endpoints are unknown; consequently, other constraints were

used in recognition. In Section 2.3, different types of speech knowledge and how each can be used as a recognition constraint are discussed. In Section 2.4, a recognition model which is based upon conclusions from the feasibility study and which uses constraint information is described. The general structure of the model is described first, followed by a discussion of how speech constraints should be used in each component of the model.

The model performs initial analysis at the broad phonetic level. Sequential constraints are applied to produce word candidates and then each word candidate is scored using more detailed phonetic analysis. The best sentence is recognized as the best scoring sequence of word candidates. The proposed model for continuous speech recognition incorporates the speech constraints described in Section 2.3 (in addition to sequential constraints) to further specify the use of speech knowledge in the Shipman and Zue model and to extend the model to continuous speech.

2.1 Sequential Constraints in Isolated Words

House and Neuburg (1977) introduced the idea of representing an utterance as a sequence of broad phonetic classes. They introduced this idea with the belief that gross linguistic categories could be identified more easily than a detailed phonetic representation. In 1982, Shipman and Zue conducted a study on sound patterns in isolated words. The results of the study demonstrated that the sound patterns of English impose strong sequential constraints on the words in the language. For example, they found that the only word in Webster's 20,000-word pocket dictionary satisfying the template:

[consonant] [consonant] [l] [vowel] [nasal] [stop]

is "splint," illustrating that sequential constraints exist even when some phonemes are represented as a broad phonetic class (relaxing the constraints). Shipman and Zue performed an experiment in which each sound was represented as one of six

broad phonetic classes: strong fricative, weak fricative, vowel or syllabic consonant, stop, nasal, and liquid or glide. The average number of words matching a particular sequence, normalized by frequency of occurrence in the Brown Corpus, was reduced to approximately 0.2% of the lexicon when this representation was used, and the maximum cohort size was about 1% of the lexicon.

These results show that strong sequential constraints exist in broad phonetic representations of words. Furthermore, since more detailed distinctions must be made to produce a detailed phonetic representation than a broad phonetic representation, a broad phonetic representation is more robust than a detailed phonetic representation. Therefore, sequential constraints at the broad phonetic level also should be more robust. Thus, Shipman and Zue proposed the following approach to isolated word recognition. The sound units are first classified into several broad categories which can be determined with little error. Then, indexing into the lexicon, a subset of the lexicon is found as word candidates. Finally, a detailed analysis of acoustic differences is used to recognize the word.

2.2 Sequential Constraints in Continuous Speech

The Shipman and Zue study demonstrated that strong sequential constraints exist on words in the English language. If strong sequential constraints were shown to exist in continuous speech, then the recognition approach outlined by Shipman and Zue could be extended to continuous speech. To test this hypothesis, a feasibility study of sequential constraints in continuous speech was conducted on a limited vocabulary, the digits from "zero" through "nine," with the idea that the task may later be expanded if the initial approach proved viable. In this section, the results of this study are outlined.

In the isolated digit vocabulary, there are two unique consonant clusters, /θr/ and /ks/, but in connected speech the word endpoints are not obvious. Thus,

consonant "clusters" can be formed from combinations of word-final consonants with word-initial consonants. In the digit vocabulary, there are 32 (ignoring gemination) unique sequences of a word-final consonant followed by a word-initial consonant. However, none of these consonant sequences are allowable within a digit. Since the set of consonant sequences at word boundaries and within a digit are mutually exclusive, all word boundary locations between two consonants in an ideal phonetic transcription can be determined by examining phoneme sequence pairs and using constraints on allowable consonant sequences.

In a broad phonetic representation, there is only one non-unique phoneme class sequence in the digit vocabulary: [stop] [strong-fricative], as in "six" and "eight seven." Thus, sequential constraints in digit strings should still exist at the broad phonetic level. This result is important because a broad phonetic representation is more easily and robustly derived automatically from the speech signal than a phonetic transcription where inter-speaker variations may be of the same magnitude as phonetic differences. In addition, such a representation may be more robust against environmental variability.

The application of detailed phonetic and broad phonetic sequential constraints were examined more qualitatively. Sequential constraints in the digit vocabulary were found to be strong enough to uniquely parse a detailed phonetic transcription of a digit string to recover the original digits. In contrast, broad phonetic sequential constraints are not as strong. An experiment was conducted to examine the application of broad sequential constraints to an ideal broad phonetic representation of digit strings for identifying individual digit boundaries from a string. The representation was ideal because word boundary effects were ignored and the broad phonetic transcription was correct. The constraints were expressed as sequences of broad phonetic classes which could represent a digit. In this study sequential constraints were used to propose words *and* corresponding word boundaries in digit strings; in contrast, in the Shipman and Zue study, sequential constraints were used

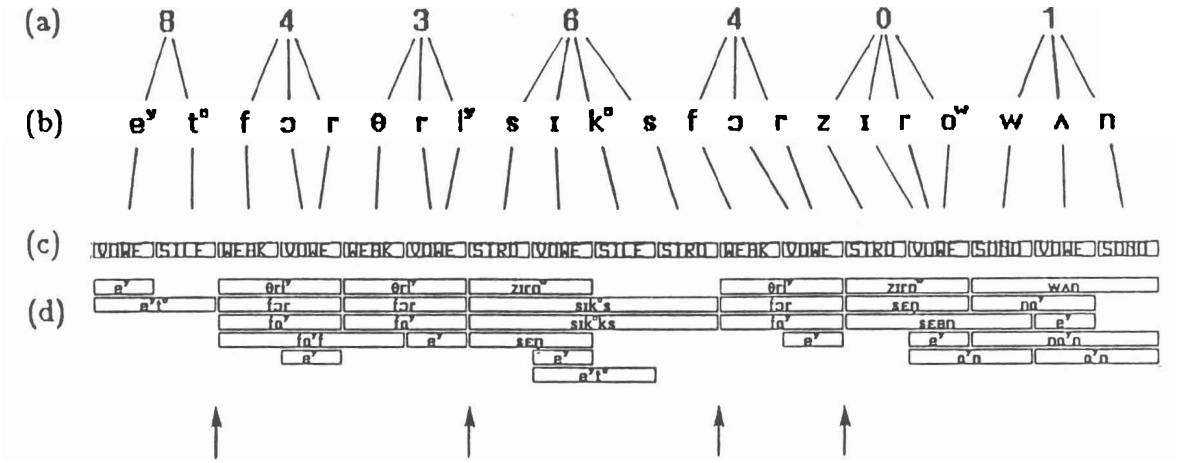


Figure 2.1: Example of Application of Sequential Constraints to an Ideal Broad Phonetic Representation

to reduce the number of cohorts associated with a word.

Ten thousand digit strings containing 34,947 word boundaries were used in the study. The digit strings were of random length and were composed of digits in random order. A phonetic representation of each string was produced by concatenating a phonetic transcription of each digit in the string. The phonetic transcription of each digit was randomly selected each time from the set of transcriptions observed for the digit in a transcribed set of training utterances, thus allowing for multiple pronunciations of a digit. The phones in the phonetic transcription were then mapped into six broad phonetic categories (strong fricative, weak fricative, silence, vowel, sonorant[1], or short voiced obstruent) to produce a broad phonetic transcription.

An example of the mapping procedure is shown in Figure 2.1. The digit string "8436401" Figure 2.1a is mapped into a phonetic string (b), and then mapped into a broad phonetic string (c). All the words with a broad phonetic representation which matches a portion of the segmentation string are shown in (d). The word boundaries marked by the vertical arrows can be identified with certainty because

[1]In this thesis *sonorant* refers to a consonant class, rather than a distinctive feature.

all word hypotheses in the region either begin or end at the segment boundary. Two word boundaries cannot definitely be identified. For example, the boundary between the first "four" and the "three" cannot be definitely identified because the word [faʸf] spans the boundary. Using this representation, 66% of the word boundaries were found. Thus strong sequential constraints exist in digit strings even at the ideal broad phonetic level.

The ability to identify 66% of the word boundaries in ideal data implies that sequential constraints can be useful in hypothesizing word candidates from a broad phonetic representation. However, these results are not directly applicable to real data. In real systems, the lexical access component must tolerate phonetic variability in the signal and front end errors. This reduces the strength of the sequential constraints and may increase the number of word candidates. To help reduce the number of word candidates, other speech constraints can be used.

2.3 Speech Knowledge for Speech Recognition

Speech can be characterized in many ways—acoustically, phonetically, by sequential ordering of sounds, and by duration of sounds. Knowledge about these speech characteristics can be used as constraints in recognition. Sequential constraints examined by Shipman and Zue were expressed using several representations, including detailed phonetic and broad phonetic representations. Similarly, other types of speech constraints can be expressed at both the phonetic and broad phonetic levels. These constraints can be used in many areas of recognition, such as segmentation and labeling the speech signal to produce a sequence of broad class labels. Different types of speech information which can be used as constraints in speech recognition at the phonetic and broad phonetic level of description are described in the following sections.

Appropriate points in the recognition process for applying each constraint are

also discussed. A constraint should be applied when it is most effective and when the system has enough knowledge to check if the criteria for the constraint to be applied are present. Before applying a constraint, the amount by which a particular constraint reduces a task should be considered. For example, when both phonetic and broad phonetic constraints provide sufficient constraint to be used, broad phonetic constraints are applied first. Constraints defined at the broad phonetic level require less detailed knowledge and are dependent upon more robust information. Furthermore, the additional information gleaned by using these constraints may allow better use of detailed phonetic knowledge.

2.3.1 Acoustic and Phonetic Knowledge

Speech scientists have developed a set of distinctive features for characterizing speech sounds (e.g., Jakobson et al., 1952). Some of these features (such as whether a vowel is high or low) have well defined acoustic correlates. Other features (such as whether a consonant is distributed) do not have obvious acoustic correlates (Fant, 1969) and therefore do not lend themselves well for use in recognition systems. Speech can also be described by a set of acoustic characteristics. Spectrograms[2] are one representation in which acoustic-phonetic information can be observed. In Figure 2.2, vowel regions are indicated by the bars underneath the spectrogram. Note the striations in the signal and presence of energy below 800 Hz in these regions. Properties of speech sounds which can be described acoustically will be defined to be *acoustic features* in this thesis. For example, a "large amount" of energy below 800 Hz satisfies this definition of an acoustic feature.

Properties of large sound classes can be described using acoustic features. For example, one property of vowels and voiced sonorants is that they are strongly voiced, and this can be described by acoustic features such as a "large amount" of

[2] A spectrogram is a frequency versus time representation of a signal where amplitude is represented by print darkness.

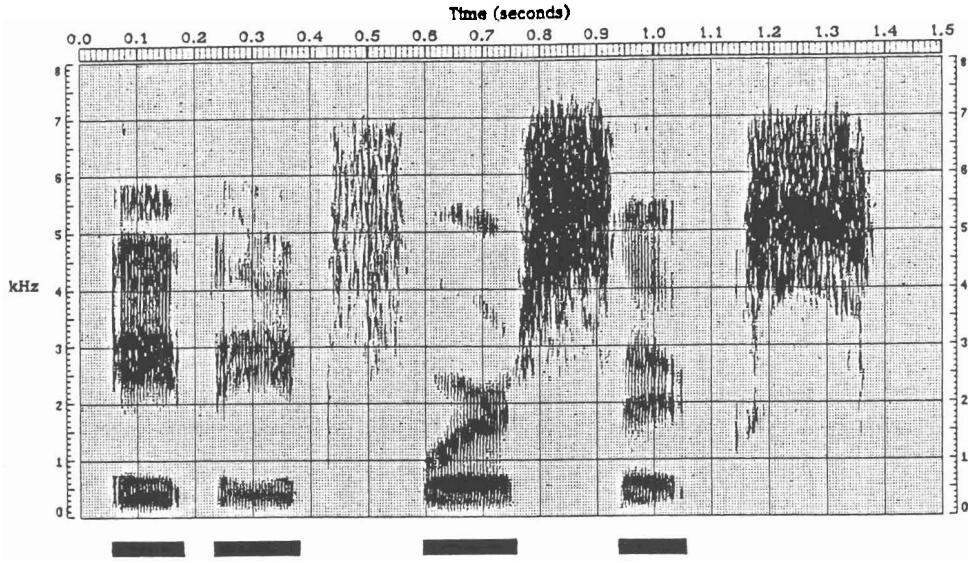


Figure 2.2: Sample Spectrogram. Bars indicate vowel regions

energy below 800 Hz. Acoustic features can also describe more detailed acoustic events, such as a rising second formant or a sharp onset. These detailed acoustic features are useful in making fine phonetic distinctions, as between a /θ/ and the release in /t/.

An acoustic feature is defined using acoustic constraints to describe a characteristic of speech. The combination of acoustic features describing a class of speech sounds define a *phonetic constraint.* For example, a strong fricative is characterized by a non-periodic energy source and a large amount of high frequency energy (an acoustic feature). The aperiodic signal, in turn, is characterized by a high zero crossing rate. Very low frequency energy may also be present in voiced strong fricatives, especially during the initial portion. The acoustic features of a large amount of high frequency energy, a high zero crossing rate, and maybe low frequency energy at the beginning of a region form one description of the strong fricative class of speech sounds and define a phonetic constraint for the strong fricative class.

Since phonetic constraints are defined by acoustic features, acoustic constraints are applied before phonetic constraints. Coarse acoustic and broad phonetic con-

straints, defining broad classes should be applied early in the recognition process because these robust descriptors provide strong constraints in narrowing down the word candidates. In addition, these constraints require little computation and no contextual knowledge. Detailed acoustic and phonetic constraints defining phones should be applied later in the recognition process. The realization of a phone is context dependent, thus how "good" a phone candidate is can be more accurately accessed given the context. Since the context is not available until at least one pass is made over the portion of the signal being considered, a recognition system does not have enough knowledge to apply detailed acoustic and phonetic constraints early in the recognition process.

2.3.2 Lexical Knowledge

Lexical information is another type of speech knowledge which can be used to constrain the recognition problem. In the English language, many patterns exist in any subset of words. As the Shipman and Zue results showed, only a small percentage of words can be represented as a particular sequence of broad phonetic classes. Thus the limit on the number of words associated with a particular sequence is one expression of lexical constraint in isolated words.

Lexical constraints are expressed in continuous speech a couple of ways. Each word in the lexicon can be represented as a sequence of broad classes or phonemes. Based on these sequential constraints, word boundaries can occur only in positions where a sequence matches a portion of the segmentation string. Once words are proposed, the correct word sequence should form a path through the lattice of proposed words. Path constraints, which are another expression of lexical constraints, can be used to prune words when at least one of the following conditions is not satisfied: 1) a preceding adjacent word exists or the word is sentence initial; 2) a following adjacent word exists or the word is sentence final. Paths which traverse these word candidates form an incomplete path in the word lattice. Pruning these

candidates reduces the computation needed in further processing.

In a recognition system, sequential constraints can be applied at the broad phonetic or detailed phonetic level to propose word candidates. Shipman and Zue's work and the results of the feasibility study discussed in Section 2.2 indicate that much constraint is available at the broad phonetic level. Since an accurate broad phonetic segmentation is more easily computed than an accurate detailed phonetic segmentation, word candidates should initially be proposed from a broad phonetic representation. Application of path constraints follows naturally, using knowledge of the endpoint locations of proposed words in the word lattice to prune word candidates which lead to a "dead end" path.

2.3.3 Knowledge of Duration

Durational constraints may be expressed in segment, word, and other representations. These constraints are derived from knowledge that the duration of a given speech unit is limited to a particular range. Durational constraints can be used to rule out a hypothesized unit which has a duration outside the observed range of the unit. For instance, a segment may represent one or more phones. In the digit string "one seven," a strong fricative segment is used to represent one phone, the /s/ in "seven." But in the phrase, "six seven," the /s/ in "seven" can share the same strong fricative segment as the final /s/ in "six"; this strong fricative segment represents two phones. In this example, durational constraints may be used to rule out the possibility that only one phone is represented by the strong fricative segment or the possibility that two phones are represented by the strong fricative segment. Figure 2.3 shows the distribution of the duration of a model segment representing one phone versus the distribution of the duration of the same model segment when it represents two phones. In region (a), only one phone is possible so a model segment with a duration in region (a) would not be allowed to represent two phones; similarly, in region (c), only two phones are possible and a model segment with a

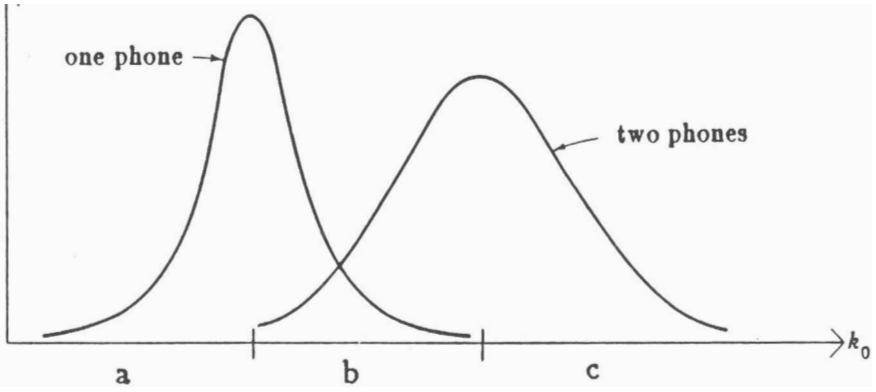


Figure 2.3: Schematized Histogram of Segment Duration for: (a) one phone (b) one or two phones (c) two phones

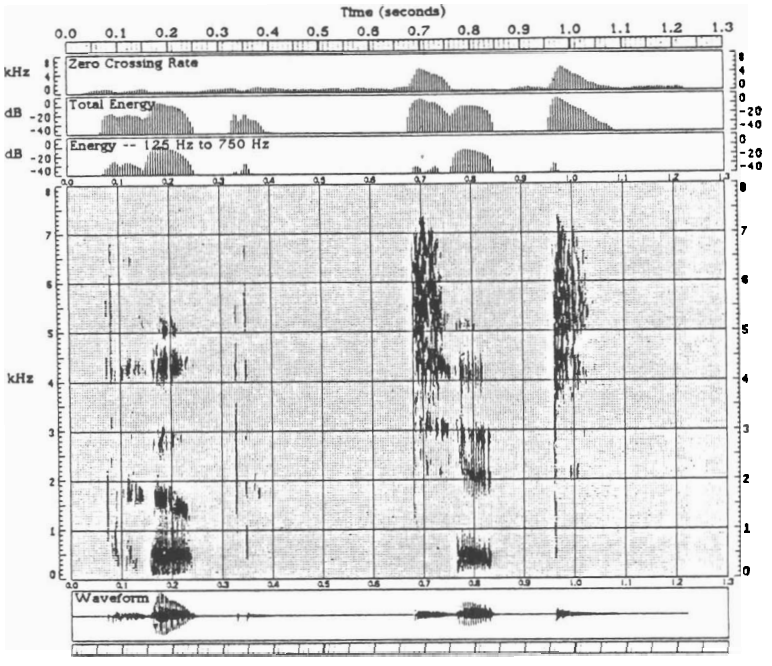
duration in region (c) would not be allowed to represent one phone; and in region (b), one or two phones are possible so neither can be ruled out.

Durations of words may also be used as a constraint. However, durations must be used with care. Speaking rate varies from sentence to sentence and also within a sentence. Consequently, unreasonable word durations must be determined using durational information from at most the sentence being recognized.

Preconditions for applying durational constraints depend only upon the presence of the unit being considered. Since durational constraints are used to *rule out* units for which the duration is unreasonable, these constraints should be applied soon after the unit is hypothesized and before further processing. For example, durational constraints on the number of phones represented by a segment can rule out some phone combinations with little computation. Detailed acoustic analysis may also be used to determine whether a segment better represents a single phone candidate or a pair of phone candidates. Since this is more computationally expensive, reducing the number of candidates by early application of durational constraints is preferred.

2.3.4 Knowledge of Allophonic Variation

Depending upon the context of a phoneme, many different realizations of that



(a) (b)

Figure 2.4: Spectrogram of /u/ in: (a) the word "poop" (labial context) and (b) the word "toot" (alveolar context)

phoneme are possible. For example, the second formant in /u/ is much lower in frequency when surrounded by labial consonants than when surrounded by alveolar consonants, as illustrated in the spectrograms of Figure 2.4. Allophonic constraints are based upon this type of knowledge. Allophonic constraints would specify that a hypothesized /u/ with a low F_2 in the context of alveolar consonants be ruled out as a candidate phoneme, but that a hypothesized /u/ with a high F_2 in the context of alveolar consonants is alright. Thus, if an /u/ is hypothesized, but F_2 is high and labial consonants are known to surround the vowel, then /u/ can be ruled out as a candidate phoneme.

Allophonic constraints can also be specified at the broad phonetic level. Different allophones of /t/ are used in pronunciations of the word "eight." Three different pronunciations of "eight" and the context, if required, for each pronunciation are shown in the first two columns of Table 2.1. Note that "eight" may be pronounced with a released or unreleased /t/ in any environment; but "eight" is pronounced with a flapped /t/ only when the following word begins with a vowel. Thus an

Table 2.1: Pronunciations of /t/

Transcription	Context	Broad Class Representation
[e^{y}t$^{\neg}$t]		vowel silence fricative[3]
[e^{y}t$^{\neg}$]		vowel silence
[e^{y}ɾ]	__vowel	vowel short-voiced-obstruent

"eight" pronounced with a released or unreleased /t/ has no contextual constraint, but an "eight" with a flapped /t/ can only be followed by a word which begins with a vowel; if none of the following words begins with a vowel, then the "eight" should be removed as a word candidate. The rightmost column of Table 2.1 shows how the three different pronunciations can also be represented using broad phonetic classes. We see that the allophonic variations which occur in /t/ can be expressed at the broad phonetic level; broad phonetic allophonic variation can also be expressed for some other consonants. Hence, allophonic knowledge at the broad phonetic level can be used to rule out word candidates when some consonant contexts are incompatible. Allophonic knowledge at the phonetic level can be used to rule out word candidates based upon more subtle differences, such as the vowel realizations illustrated by the earlier /u/ example.

As previously stated, broad phonetic constraints should be used before phonetic constraints when both types of constraints are used. Thus, allophonic knowledge at the broad phonetic level is used first to prune the word lattice when a broad phonetic representation of each word candidate is available. Allophonic constraints at a detailed phonetic level are used later to discriminate between word candidates based upon fine differences in similar speech sounds.

[3]The release of a /t/ actually consists of a burst followed by aspiration. In this thesis, the broad class of "fricative" was generalized to include aspiration in addition to fricative sounds.

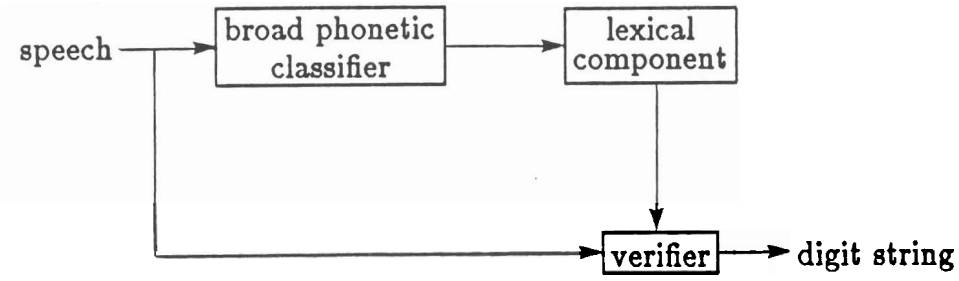


Figure 2.5: Phonetically-based continuous digit recognition system

2.4 Recognition Model

In this section, an acoustic-phonetic continuous speech recognition model is developed. The model's particular configuration was motivated by the results of the feasibility study described in Section 2.2 and the philosophy of using speech knowledge to constrain the recognition task.

In Shipman and Zue's proposed model for isolated word recognition, the speech signal is represented as a sequence of broad classes from which word candidates are hypothesized. Each word candidate is then analyzed in more detail and the word is verified or rejected. This same idea can be used for continuous speech. From a broad class representation of a sentence, word candidates and also the corresponding endpoints are hypothesized, producing a lattice of words. Each word candidate in the lattice is analyzed in more detail and then scored. Finally, the best sequence of words spanning the lattice is found.

The main components of this system, shown in Figure 2.5, are a broad phonetic classifier, lexical component, and verifier. The broad phonetic classifier produces a broad class segmentation of the incoming signal based on coarse characteristics of the parameterized signal. The use of coarse characteristics by this component makes its output potentially robust with regard to speaker variability. The broad class segmentation is then input into the lexical component where the phonetic tran-

scription of each word in the lexicon is matched against the segmentation produced by the system. This yields a lattice of word candidates; the lattice is further reduced using lexical, allophonic, durational, and contextual constraints. When more than one word candidate exists over a portion of the reduced lattice, the verifier will rank order the candidates based on detailed acoustic analysis and knowledge about feature characteristics of the phones in each word candidate.

Speech constraints are applied at appropriate points in processing of the system. The most general constraints are applied first, with each new constraint being more specific. The constraints used by the broad phonetic classifier to produce a broad phonetic representation are: coarse acoustic features of speech sounds, broad phonetic characteristics, and durational constraints on broad phonetic classes. The broad phonetic classifier first derives coarse acoustic features from the speech signal. Broad phonetic classes are hypothesized based upon the coarse acoustic features present and broad phonetic constraints. Once broad phonetic classes are hypothesized, durational constraints are applied to check that the duration of a segment is reasonable. For example, a short voiced obstruent, such as an intervocalic /v/, should be of shorter duration than most of the other segments.

The constraints relevant to producing a lattice of word hypotheses are sequential constraints, durational constraints, path constraints, and broad allophonic constraints. The lexical access component first applies sequential constraints to produce word hypotheses with associated time endpoints. Because many word candidates are hypothesized, other constraints are useful in reducing the number of candidates to a manageable number for verification. Durational constraints dictate whether a segment must represent one or two phones. For example, when duration constraints dictate that a segment represents only one phone, then all paths which require the segment to represent two phones can be removed. Path constraints, which do not require any extra processing before they can be used, are applied next to rule out word candidates which would form an incomplete path. Broad allophonic con-

straints are used to remove words from the lattice with context requirements which are incompatible with all the previous or following words.

Verification uses knowledge about detailed acoustic features of speech and detailed phonetic characteristics. Each phone is scored based on how well the acoustic features of the unknown segment match the expected acoustic feature values dictated by the phonetic characteristics of the hypothesized phone.

Thus the model uses knowledge which can be derived robustly to constrain the task at each stage of processing. Broad phonetic constraints are applied first, followed by more directed use of detailed constraints. Although use of only low-level speech knowledge is addressed, higher level knowledge can be incorporated into the model in the verification component.

2.5 Chapter Summary

In this chapter, the background leading to a model for continuous speech recognition was developed. The main issues discussed were:

- Shipman and Zue showed that strong sequential constraints exist on the words in English at the broad phonetic level. Based on this result, they proposed an isolated word recognition model.
- Strong sequential constraints also exist in continuous digits; therefore, the Shipman and Zue model can be extended to continuous digits.
- Many different types of low-level speech knowledge may be applied to the recognition task.
- A model for continuous speech recognition based upon the use of broad phonetic sequential constraints was proposed. Other types of speech knowledge were also incorporated into the model.

Chapter 3

Broad Phonetic Classification and Lexical Access

Speech from the same speaker saying the same phrase is never identical, and speech from different speakers contains even greater differences. These interspeaker and intraspeaker differences in speech occur because natural speech is not a sequence of discrete units; it is a continuum of sounds. Speech sounds are a function of the jaw and tongue position, and the continuous movements of the articulators modify the vocal tract configuration to produce a time varying signal. Depending upon the speaking rate and the current configuration of the articulators, the "target" for the next sound is reached with different degrees of accuracy before movement begins toward the configuration of the following sound.

Utterances of the same sentence vary in rate, pronunciation of each word, and degree of coarticulation. Speaking rate is affected by factors such as a speaker's mood and to whom the speaker is talking (e.g., child or adult). As the speech rate varies, the durations of different speech sounds vary nonlinearly; for example, as the speech rate becomes slower, vowel duration increases much more than stop burst duration. A speaker may pronounce the same phoneme using different realizations because of phonetic context, as in the example with /u/ in alveolar and labial

contexts, or for other reasons. For example, a person may sometimes say "eight" with a released /t/ and sometimes with an unreleased /t/. A speaker may even delete sounds, such as pronouncing "eight six" as [eʸsɪk¬ks], where the /t/ in "eight" is deleted, or insert sounds, such as pronouncing "three" as /θəri/ where a /ə/ has been inserted. In addition, different acoustic realizations of a phoneme may be used by different speakers. For example, speakers may pronounce an intervocalic /v/ as a canonical voiced weak fricative, an unvoiced weak fricative, or the /v/ may be so weak that it approaches silence.

A speech recognition system must handle these variables. The system must know about common factors for each sound and/or know about the causes of variation and use this knowledge in recognition. In our implementation of the continuous speech recognition model, variability in natural speech was handled at two levels. The broad phonetic classifier dealt with variability within a class of sounds, and the lexical access component dealt with segmentation errors due to variability in the broad class representation of a sound.

A broad phonetic representation must be at least as accurate as a detailed phonetic representation because a broad phonetic representation is a description that is embedded within a detailed phonetic description. Thus, less detailed distinctions need to be made to produce a broad phonetic representation than a detailed phonetic transcription. In a broad phonetic representation, many speech variabilities, such as whether an /u/ is fronted are of no significance. The phonetic classifier labels an /u/ as a vowel whether or not it is fronted. Thus by knowing what characterizes different broad classes of speech sounds, the broad phonetic classifier labels speech at a broad phonetic level much more accurately than it could label speech at a detailed phonetic level.

However, segmentation errors can still happen at the broad phonetic level and unanticipated acoustic realizations can still occur. For instance, incomplete closure may occur in a stop gap, resulting in a "noisy" stop gap (see Figure 3.1) which

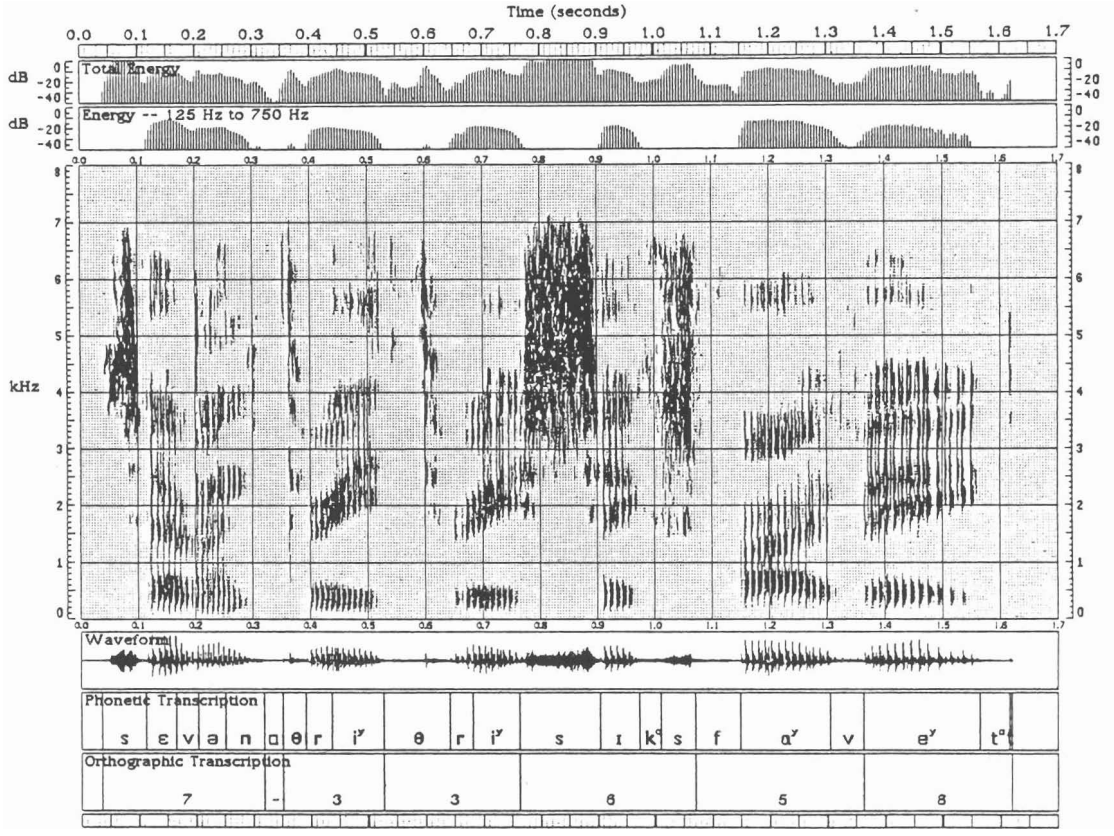


Figure 3.1: Noisy stop gap in "six" in digit string "733658"

may be labeled as a weak fricative. The lexical access component handles these segmentation errors due to speech variability using two types of knowledge: 1) how often a phoneme is mislabeled as another class and 2) how often a phoneme is labeled as a particular class given its context. An example of the first type of knowledge is how often a /k/ closure is labeled a "weak-fricative" instead of "silence" by the broad phonetic classifier. The second type of knowledge includes knowledge about when insertion or deletion errors occur. Examples of this type of knowledge include how frequently the /n/ and /s/ in the sequence /ns/ are both labeled as "strong-fricative" (a deletion) and how frequently an /n/ is labeled as a "sonorant" when preceded by an /s/ which was labeled a "strong-fricative" (a match). Together the broad phonetic classifier and lexical components take into account many of the

variabilities in speech to produce viable word candidates.

3.1 Broad Phonetic Classification

The broad phonetic classifier segments the speech signal and labels each segment as either silence or a broad phonetic class: strong fricative, weak fricative, short voiced obstruent, sonorant, or vowel. These classes were chosen because they could be robustly identified and correspond to different manners of articulation.

Classification is done by parameterizing the speech signal, extracting acoustic features from the parameters, and labeling segments based upon the acoustic features present. In an effort to minimize labeling errors, robust information as well as delayed binding was used in the classification approach. By allowing multiple segment labels until the final stage in processing, the classifier can use all the information learned from earlier stages in processing to make a final decision on labeling the segments.

3.1.1 Parameterization

Parameters were computed from speech digitized at 16 kHz and lowpass filtered at 6.4 kHz. The sampling rate was chosen to include the frequencies containing most of the speech information. Disagreement exists about the frequency range of speech: Hyde (1972) stated that speech "covers a frequency range of about 10 kHz"; in contrast, telephone speech has a passband of 300 Hz to 3300 Hz and is acceptably intelligible–however, this may be due to use of higher level constraints by listeners. In this study, the overriding consideration in choosing a sampling rate is the fact that Zue and his students can read spectrograms computed from speech sampled at 16 kHz and filtered at 6.4 kHz, indicating that enough information is present in the speech signal to be recognized when processed in this way. The information in the higher frequency range is important in identifying phonemes with energy

concentrations in the higher frequencies (e.g., /s/). This is especially true in the analysis of female speech, which has higher natural frequencies due to the shorter average vocal tract length of females.

Acoustic parameters were defined for representing the speech signal in a compact format. Based on spectrographic examination of digit strings, candidate parameters which appeared to capture robustly the occurrence of significant events in a spectrogram were designed. The parameters in the final set were chosen for their usefulness in identifying and differentiating among different classes and for their robustness. The chosen parameters are energy in various frequency bands and zero crossing rate.

In most of the parameter computations, pre-emphasized speech was used since pre-emphasis compensates for the spectral tilt of the speech spectral envelope. This gives the higher frequencies, which are predominant in sounds such as /s/ and /θ/, approximately equal weight to the lower frequencies. The speech waveform was normalized before any computations so that the maximum of the sample values in each utterance is the same. The energies were calculated as log energy to minimize sensitivity to small variations when the energy values are large. To avoid rapid changes in value as a formant moved in or out of an energy band, tapered frequency windows were used to compute the energy within a frequency band from the DFT. The DFT's were computed every 5 msec using a 25.6 msec Hamming window. A sample tapered frequency window is shown Figure 3.2 and the corresponding frequency points used in the energy calculations are shown in Table 3.1.

A spectrogram of the digit string "6861994" is shown in Figure 3.3b, and corresponding parameters used in coarse acoustic analysis are shown in Figure 3.3a. Note that the low-frequency energy (energy 125-750 Hz) is highest in vowel and sonorant regions. This is because F_1, (and possibly a nasal formant) is present during the production of vowels and voiced sonorants. Thus low-frequency energy is a good indicator of voiced regions.

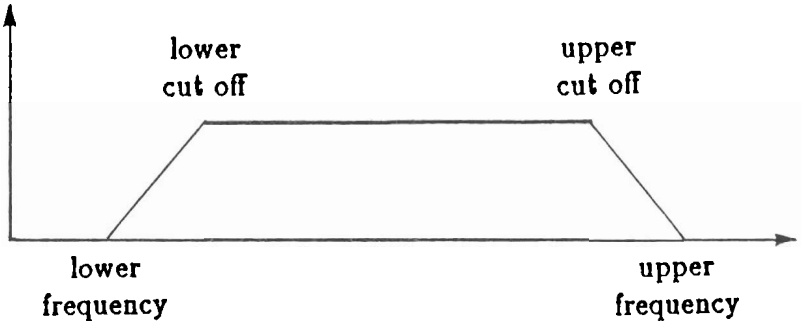


Figure 3.2: Sample Tapered Frequency Window

Table 3.1: Tapered Energies Used in Coarse Acoustic Analysis

Parameter	lower frequency	lower cut off	upper cut off	upper frequency
energy 125-750 Hz	0	125	750	900
energy 1000-2000 Hz	800	1000	2000	2200
energy 1000-3000 Hz	800	1000	3000	3200
energy 4500-7800 Hz	4300	4500	7800	8000

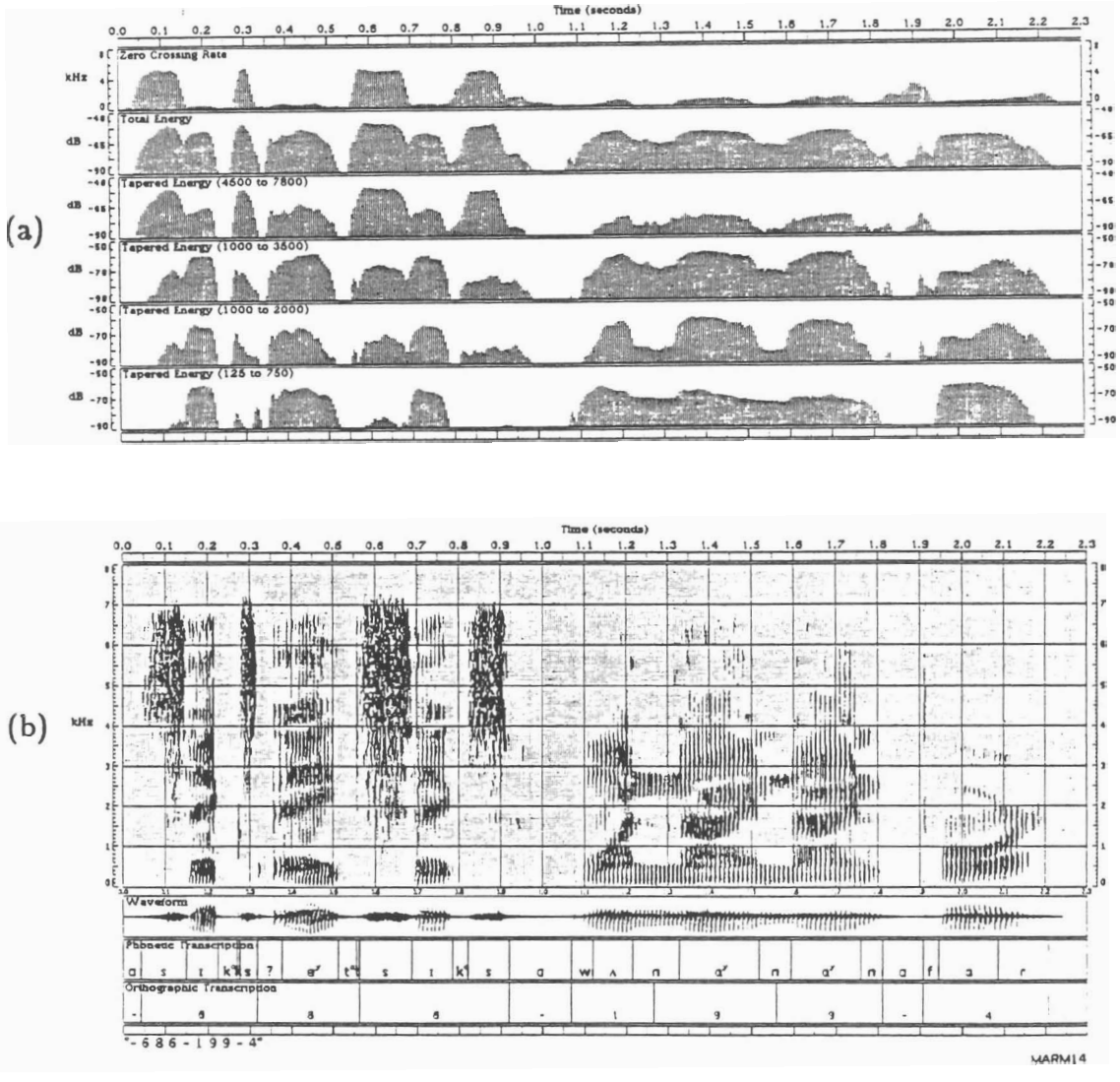


Figure 3.3: The Digit String "6861994" (a) Sample Parameters and (b) Spectrogram

Intervocalic /n/'s have been observed to dip rapidly in the lower-mid-frequency energy (energy 1000-2000 Hz). The dip is due to a nasal zero in that frequency region and the abruptness of the dip is due to the output quickly switching from the oral to nasal cavity. Note in Figure 3.3 the dip in lower-mid-frequency energy during the intervocalic /n/ of the "one nine" and "nine nine" sequences at around 1.25 and 1.55 sec, respectively. Other energy parameters show a slight dip during production of the intervocalic /n/, but the lower-mid-frequency energy dip is much more robust. This example stresses the importance in selecting appropriate parameters to characterize an acoustic feature.

Total energy is lowest during pauses and stop gaps (e.g., at 1.0 and .24 sec in Figure 3.3); it is usually lower in weak fricatives than strong fricatives (compare total energy in the /s/ at 0.10 and the /f/ at 1.92 sec). Thus total energy can be useful in indicating silence and in discriminating between weak and strong fricatives.

High-frequency energy (energy 4500-7800 Hz) is largest in the presence of strong fricatives. The figures show that fricatives, especially /s/ and /z/, generally have more energy in the higher frequencies than vowels. This is because vowels have formants beginning at about 300 Hz, and each formant "tilts" the spectrum above it down by 12 dB/octave (Fant, 1970). In contrast, the lower frequency poles in fricatives are canceled by zeros; consequently the energy in fricatives is concentrated in the higher frequencies.

Strong vowel formants are usually observed in a spectrogram up to at least 3000 Hz, and even higher for front vowels. In many nonsonorants, like /v/, there is little energy in this region. The mid-frequency energy parameter (energy 1000-3000 Hz) tries to capture the contrasting presence of mid-frequency energy in vowels and lack of it in consonants. In some consonants there is significant energy in the mid-frequency range, such as a rounded /t/ in "two," but in these cases, low-frequency energy, which is usually weak during voiceless consonants like /t/, may be used as a secondary cue to rule out the candidate as a vowel.

The zero crossing rate is usually high during the production of fricatives (time 0.10 sec) and aspiration, because of the turbulence associated with the production of these sounds. The zero crossing rate is generally much lower in vowels, although the amount depends primarily upon the amount of high-frequency energy present in the speech of a speaker.

3.1.2 Acoustic Feature Extraction

During an utterance, parameters, such as the ones described above, exhibit salient features corresponding to speech sounds. For example, the zero crossing rate is high during voiceless fricatives. Algorithms for extracting a set of these *acoustic* features, as defined in Chapter 2, were designed. The feature set was composed of the descriptors *high*, *low*, *dip*, and *rapid transition*. *High* and *low* (high-low features) indicate the value of parameter in a region relative to values over the whole utterance and a set of standard value ranges. *Dip* indicates a region of lower value within a region classified as high. *Rapid transition* indicates that a region or dip has a rapid onset or offset. Not all parameter descriptors were computed for each parameter; instead only those descriptors which are robust indicators of a class or several classes of speech sounds were computed.

Broad classes of speech sounds have relatively stable characteristics during the middle portion of a segment. For example, vowels exhibit voicing. In contrast, the characteristics may change at different times during transitions between two broad classes of sounds. For example, in the transition from a vowel to a fricative, energy in the higher formants may weaken sooner than energy in the first formant. To avoid forcing a decision at each sample, which would result in less certain decisions in transition regions, the high-low descriptors label only robust *regions* where a parameter is relatively stable, and other regions are left unlabeled.

The high-low regions were found using an algorithm which depends on two thresholds, T1 and T2, to locate a region and then define the edges of a region. By

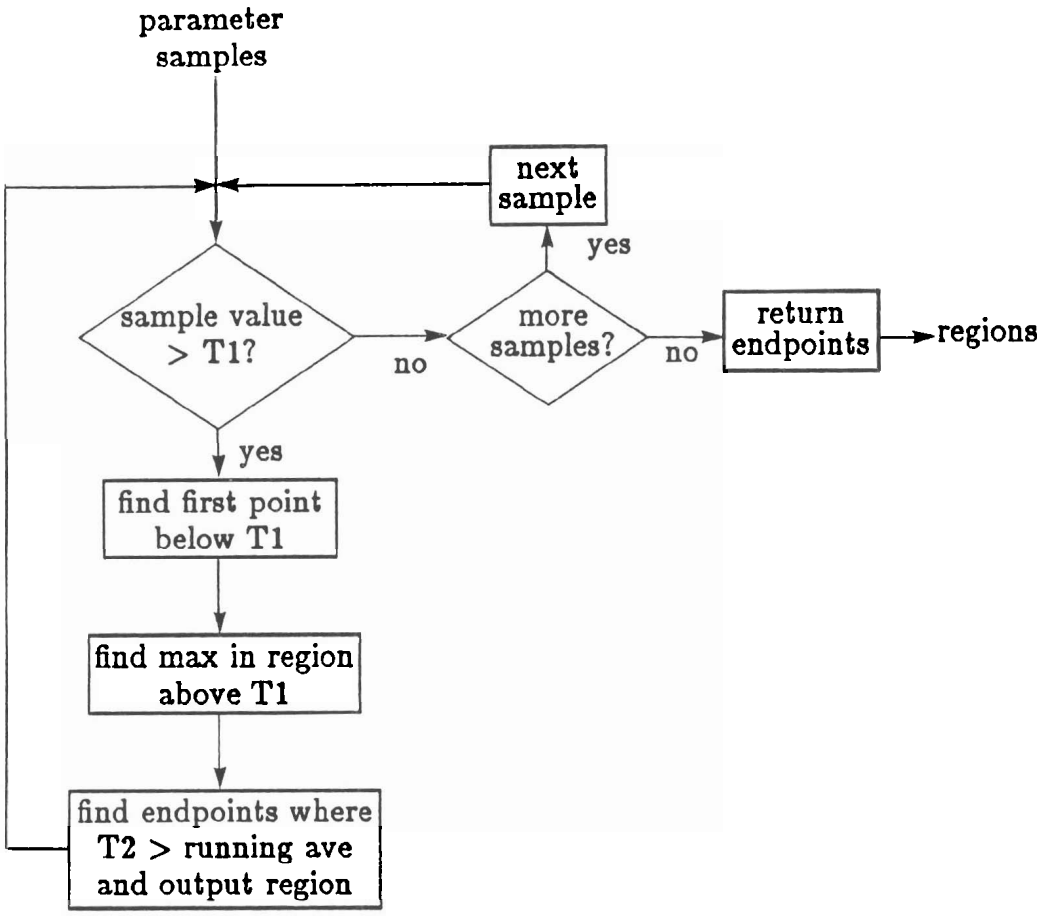


Figure 3.4: Flowchart for Finding "high" Regions

using two thresholds, robust regions, islands of reliability (Woods, 1981), can first be identified, and then anchoring from each robust region, the edges of the region can be extended. The high-low acoustic features were defined for each parameter by choosing different values of T1 and T2.

A flowchart for finding peaks using the high-low algorithm is shown in Figure 3.4. To locate high regions, points where a parameter value is greater than T1 are found first. T1 is dependent upon the minimum and maximum values observed in the utterance to be recognized and a "standard" set of values derived from a set of training utterances. T1 was defined as:

$$T1 = c_1(max_u - min_u) + min_u$$

where the max for the utterance, max_u is:

$$max_u = \begin{cases} max_{observed} & max_{observed} > max_{global} - r(max_{global} - min_{global}) \\ max_{global} & otherwise \end{cases}$$

and the min for the utterance, min_u, is:

$$min_u = \begin{cases} min_{observed} & min_{observed} < min_{global} + r(max_{global} - min_{global}) \\ min_{global} & otherwise \end{cases}$$

Max_u and min_u provide adjustment of threshold values for each utterance, allowing some adaptation to different speakers and/or environment. The constant r was chosen empirically to be .3 and is used to specify the range of values of $max_{observed}$ for which max_u is set to $max_{observed}$.

The condition on $max_{observed}$ and $min_{observed}$ insures that max_u and min_u are set to observed values for the utterance only if reasonable $max_{observed}$ and $min_{observed}$ values were computed. When the maximum(minimum) value of a parameter in the input utterance was within r of the "standard" maximum(minimum), the observed maximum(minimum) was considered reasonable and the sentence was assumed to contain at least one phone for which the parameter usually reaches the maximum(minimum) value. Conditionally adjusting the threshold in this way prevents errors such as lowering the threshold which defines regions of high zero crossing rate when no fricatives or stops are present in the utterance.

Once a region has been located, its edges are found using T2. The parameter is smoothed locally from the peak using a running average to minimize local perturbations:

$$s[i] = c_2 x[i] + (1 - c_2) s[i - 1]$$

where $s[i]$ is the value of the smoothed parameter i samples away from the peak (call the peak value max_{local}), $x[i]$ is the original parameter and c_2 is a constant chosen to be .25. The endpoints of a high region are defined to be the time when the value of the smoothed parameter has fallen to a predefined fraction (c_3) of the

difference in value between the peak in the region and min_u. Thus T2 is defined as:

$$T2 = c_3(max_{local} - min_u) + min_u$$

The use of T2 allows the endpoints of a region to be set relative to the peak value of the region and independently of T1. Additionally, if the parameter is noisy and several alternating samples dip below T1, one region is found, rather than separate regions. *Low* regions are found using the same algorithm, but with complementary thresholds and comparators.

Dips for a parameter are found only in *high* regions using a simple dip detector. The slope is computed on the median smoothed parameter, and candidate dips are hypothesized at the point that the slope changes from negative to positive (call it P_1). From this point a local max in the smoothed parameter is found on each side. If the minimum dip depth, defined to be the minimum of the difference between the parameter value at each local max and P_1, is greater than 5 dB in the smoothed parameter, then the region between the two local maxima is defined to be a dip. Dips were computed on a 3-point median smoothed parameter and 7-point median smoothed parameter. The two different smoothers were used to capture short dips (3-point) and longer dips (7-point). The dips found on the 3-point and 7-point smoothed parameters were combined to form the list of dips for the parameter.

The transition rate is checked at edges of regions or dips associated with the energy from 1000 to 2000 Hz. For each edge, the maximum slope within 20 msec of the edge is found. If the maximum slope is greater than 1 dB/msec, then the transition is labeled "rapid."

To help moderate the effect of noise, redundant features can be used to indicate the presence of a phonetic class. Redundant features were in the same spirit in which Otten (1971) hypothesized that man utilizes redundancies in noisy situations.

A second algorithm for finding when the low-frequency energy is high was used. Rather than looking for regions where the parameter values are large, the second algorithm looks for edges using the fact that onsets in voiced sonorant regions

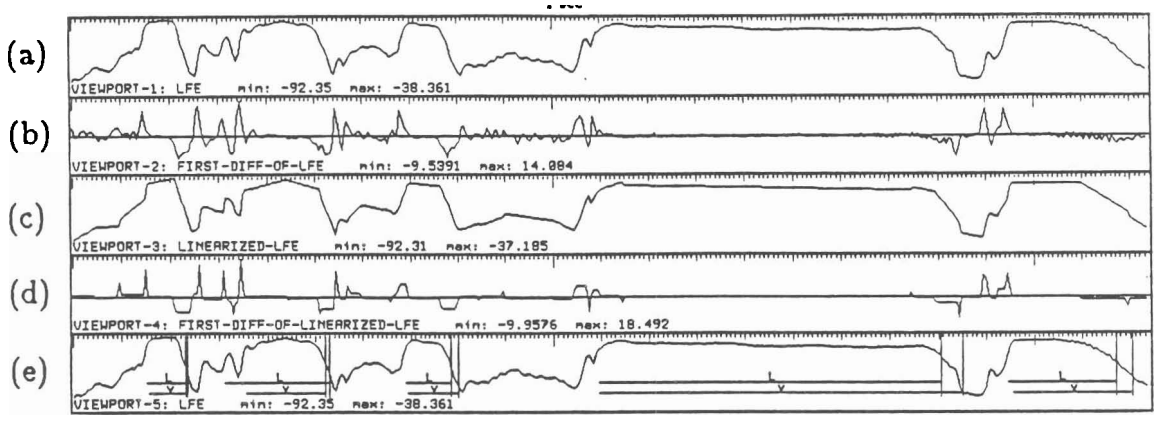


Figure 3.5: Low-Frequency Energy Characterizations of the Digit String "6861994" (a) low-frequency energy contour (b) first difference of the low-frequency energy contour (c) piecewise linear approximation to the low-frequency energy contour (d) first difference of piecewise linear approximation to the low-frequency energy contour (e) regions where low-frequency energy was found to be "high" using thresholding method "L" and edge detection method "v"

are generally characterized by a sharp rise in low-frequency energy and offsets are characterized by a more gradual decline in low-frequency energy. Thus onsets are detected first, and then an offset is found between each pair of onsets.

Figure 3.5 depicts parameters used in edge detection. The edge detector method locates onsets and offsets (edges) in the low-frequency energy contour (Figure 3.5a) based upon the first difference of a piecewise linear approximation to low-frequency energy (Figure 3.5d). A piecewise linear approximation (Pavlidis, 1974) was chosen for smoothing because it preserves the edges of the contour while smoothing small irregularities (compare (a) and (c)), resulting in a cleaner first difference of the linearized parameter (d). Because onsets in the low-frequency energy contour are sharper and can be more robustly detected than offsets, they are located where the

first difference of the linearized signal is large and the energy is above a minimal threshold. The energy threshold prevents false or early triggering, which happens when the spectral distribution of the energy in fricatives dips low. Between each pair of onsets, there must be an offset; the offset is found where the linearized first difference is most negative.

The output of the edge detector is shown in Figure 3.5e and regions between an onset and offset are indicated by a "v." Horizontal lines indicate the times the detector found the low-frequency energy to be high. The leftmost edge of a horizontal line marks the left endpoint and a vertical bar marks the right endpoint of the detected feature. The output of a threshold detector is also shown in Figure 3.5e; the regions in which the low-frequency energy was above a threshold are marked by an "L." Comparison of the two methods show close agreement in most cases.

When both detectors indicate that the low-frequency energy in a a region is high, then there is strong evidence that this is a voiced region. When there is disagreement between the two detectors, the low-frequency energy is not obviously high, but some acoustic event has occurred which causes the low-frequency energy to not be low. For example, this may be a stop with a release which extends into the low-frequency range. Thus by combining the output from both detectors, better decisions may be made.

Output from the feature detectors (rate information is not shown) for the utterance "6861994" is shown in Figure 3.6. The key to the symbols used in Figure 3.6 is given in Table 3.2. Note that most feature detectors turn on in a consistent manner: "h" is on during strong fricatives, "L", and "v" are on during voicing, and "D" is on during intervocalic nasals.

3.1.3 Broad Phonetic Labeling

Broad phonetic labeling uses a set of production rules to deduce possible broad classes from the chosen set of acoustic features. The hypothesized broad classes

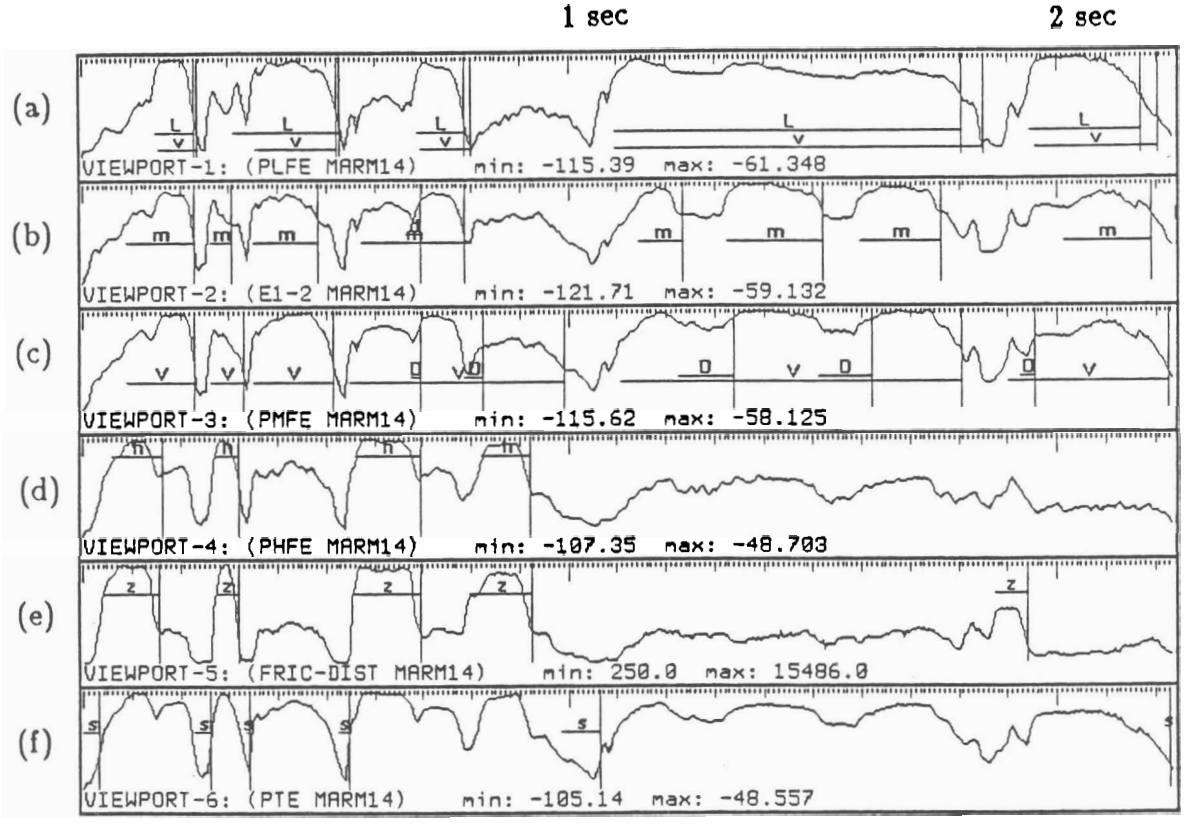


Figure 3.6: Feature Detectors for the Digit String "6861994" (a) energy 125-750 Hz is high using threshold "L" or edge detector method "v" (b) energy 1000-2000 Hz is high "m" or a dip occurs "d" (c) energy 1000-3000 Hz is high "V" or a dip occurs "D" (d) energy 4500-7800 Hz is high "h" (e) zero crossing rate is high "z" (f) total energy is low "s"

Table 3.2: Key to Figure 3.6

Symbol	Feature
h	high energy 4500-7800 Hz
z	high zero crossing rate
s	low total energy
m	high energy 1000-2000 Hz
d	dip in high energy 1000-2000
V	high energy 1000-3000 Hz
D	dip in high energy 1000-3000
l	high energy 125-750 Hz
v	high energy 125-750 Hz (edge-method)

form a segment lattice. When the production rules have deduced all candidate labels for each segment in the segment lattice, the computed values of the acoustic features are used to find the best label for each segment, thus producing a unique segmentation string. By first finding all possible labels, more directed analysis may be performed to reduce the segment lattice to a segmentation string by using knowledge of each competitor.

The production rules are applied in levels to produce the segment lattice. (See Appendix B for sample production rules.) By using multiple levels of rules, the knowledge gained by applying rules at lower levels can be used by higher level rules. Thus a rule can use contextual constraints requiring the presence of a preceding vowel if rules hypothesizing vowel-like segments have previously been applied.

The first set of 12 production rules hypothesizes each segment to be zero or more *phone-like* classes, based upon the presence or absence of combinations of non-conflicting robust acoustic features characterizing each segment. The acoustic features used are shown in the top of Table 3.3, and the phone-like classes used are

shown in the center of Table 3.3.

The second set of 16 production rules hypothesizes *phone* classes from the phone-like classes using durational and contextual constraints to rule out some of the hypothesized phone-like classes. The duration of acoustic features is useful in classification; for example, an intervocalic, short, voiced obstruent is allowed a maximum duration of 55 msec and must be preceded and followed by a vowel. Contextual constraints are used to check that the context of a given label is correct. For example, an intervocalic, short voiced obstruent must be preceded and followed by a vowel. In addition, some speech sounds have slightly different realizations depending upon context. Intervocalic nasals are characterized by a sharp dip in lower-mid-frequency energy, but in non-intervocalic nasals, lower-mid-frequency energy may "fade out." To capture these differences within a class, a separate rule is used to describe each realization and its context. The phone classes used are shown in the lower portion of Table 3.3.

Figures 3.7a and b show the segment lattice produced by the classifier after the first and second sets of production rules are applied. The key to the symbols used in Figure 3.7 is given in Table 3.3. The segment from 0.23 to 0.26 sec is labeled as silence. It can be observed that silence segments are characterized by a low amount of total energy. Similarly, strong fricatives are characterized by a high zero crossing rate and a large amount of high-frequency energy, as in the segment from 0.82 to 0.92 sec, which is labeled as a strong fricative.

Each segment is not necessarily labeled as a broad phonetic class after the first two levels of production rules are applied. If there is conflicting information such that the cues are not robust enough to make a good decision (as in a transition), the segment is left unlabeled. Short unlabeled segments less than 40 msec in duration are handled first by arbitrarily splitting each evenly between the two adjacent segments on the assumption that they represent transitional regions. This produces a segment lattice without transitions and frees the lexical component from

Table 3.3: Key to Symbols Used in Figure 3.7

	Symbol	Description
acoustic features	h	high energy 4500-7800 Hz
	z	high zero crossing rate
	s	low total energy
	q	rapid transition in energy 1000-2000 Hz
	d	dip in high energy 1000-2000 Hz
	m	high energy 1000-2000 Hz
	D	dip in high energy 1000-3000 Hz
	V	high energy 1000-3000 Hz
	L	high energy 125-750 Hz
	v	high energy 125-750 Hz (edge-method)
phone-like classes	Fl	strong fricative like
	Wl	weak fricative like
	Sl	silence like
	Vl	short voiced obstruent like
	Rl	sonorant like
	VWl	vowel like
phone classes	F	strong fricative
	W	weak fricative aspiration
	S	silence
	V	short voiced obstruent
	R	sonorant
	VW	voiced sonorant, vowel

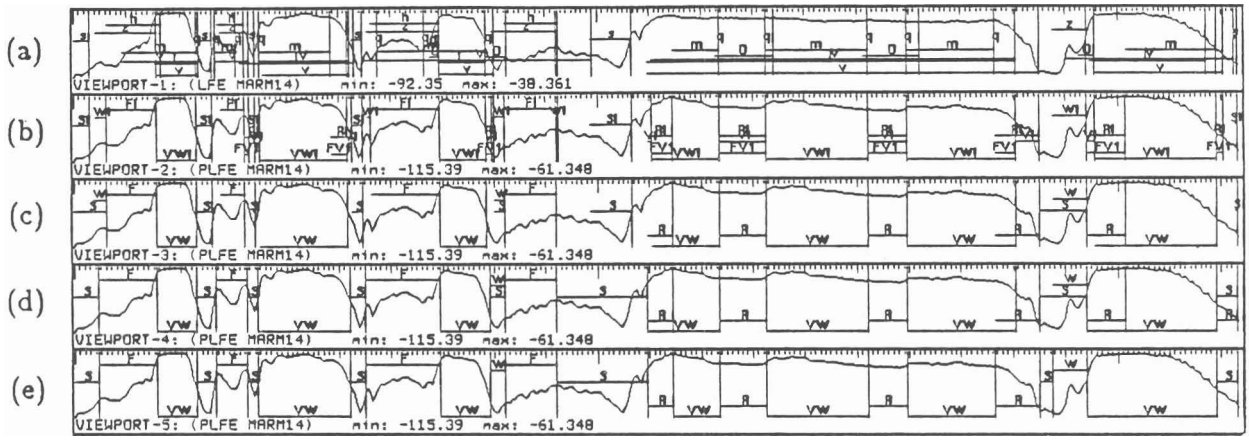


Figure 3.7: Steps in Broad Phonetic Labeling from Acoustic Features for the Digit String "6861994"

containing a subroutine for handling transition regions. In lexical access, boundary locations are irrelevant, only the order and approximate duration of the segments are important. In verification, the transition regions (boundaries) may be used to help identify a segment, but the central portion of the segment also contains much information which can be used for segment identification. Thus the philosophy was adopted that only the approximate location of boundaries need to be identified for recognition.

Unlabeled segments potentially match all broad classes and reflect no information. In these segments a consistent set of features was not present. A new feature set, called *not-features* are defined by relaxing the constraints on the high and low definitions. The not-features delimit regions where parameter values are "not high" or "not low" and are computed using the "high-low" algorithm. The set of features together with durational constraints are used to hypothesize possible label candidates and rule out definitely incorrect labels in the segments. In Figure 3.7c there

is an unlabeled segment beginning at time 0.92 sec. A corresponding segment lattice illustrating all labels after rules using "not-feature" values are applied is shown in Figure 3.7d; we can observe that the unlabeled segment beginning at time 0.92 sec was labeled as silence since all other possibilities have been ruled out by the not-features.

Once all broad phonetic labels are hypothesized for each segment, the segment lattice is reduced to a unique segmentation. Any segment which has been assigned more than one label is examined in more detail to determine the best label. Six acoustic features derived from the initial parameter set are used to capture parameter characteristics similar to the acoustic features used for initial segmentation. Since segment regions have been defined at this point in processing, computation can be performed over a specified region. Thus, rather than looking for a "high" region in an acoustic parameter, the maximum value of an acoustic parameter within a specified region is computed. The six acoustic features are: maximum pre-emphasized total energy in the center region, maximum pre-emphasized low-frequency energy in the center region, minimum total energy in the center region, minimum pre-emphasized lower-mid-frequency energy, maximum pre-emphasized mid-frequency energy, and maximum zero crossing rate. The center region, defined as the region from the quarter point to three-quarter point of the segment, was used to minimize transition effects on the computed values.

The label for each segment is chosen to be the label which is "most likely"; the likelihood of the label is computed based upon the distribution of the six acoustic features observed for each label and the value of each acoustic feature over the segment. In particular, the likelihood of label i, L_i, is computed as the product of the likelihood of label i versus label j, L_{ij}, over all j competitor labels:

$$L_i = \prod_{j \neq i} L_{ij} \tag{3.1}$$

The likelihood ratio of labels i and j is defined to be the product of the likelihood

ratio of label i versus label j based on feature f, L_{ijf}, over the six acoustic features:

$$L_{ij} = \prod_f L_{ijf}$$

Thus the more likely each feature indicates that the label is label i, the more likely the segment is label i (rather than label j). Note from Equation 3.1, that the more likely a label is relative to each competitor label, the more likely it is that the segment is that particular label.

For each pair of competitor labels, i and j, L_{ijf} is computed from the value of the feature in the segment. The probability of a label given the observed feature value is estimated using k-nearest-neighbor estimation (Duda and Hart, 1973). The probability estimate for label l and feature f, P_{lf}, is used to compute the likelihood ratio between the labels i and j for feature f:

$$L_{ijf} = P_{jf}/P_{if}$$

Since a label is more likely the larger P_{lf} is, label l is more likely the larger L_{ijf} is. This algorithm was applied to the segment lattice to produce the final segmentation string shown in Figure 3.7e.

3.1.4 Evaluation

The output of the broad phonetic component was evaluated for insertions, deletions, substitutions, and matches when compared to a hand-labeled phonetic transcription (see Appendix A for a description of the phonetic labeling procedure). The broad class transcription of each utterance was derived by mapping each phone in the hand transcription to a broad class. The broad class hand transcription and automatic broad class transcription were aligned using a simple 50% overlap criterion: if segment A_i in string A covered over half the duration of segment B_j in string B, then segment A_i was mapped with segment B_j. An overlap criterion, rather than a string alignment was chosen because the time boundaries associated with

the segment regions are used later in verification. Since correct time boundaries are relevant, a string alignment is a less stringent criteria because matches are allowed even though the boundaries are shifted.

Insertions were defined as two adjacent segments from the automatic transcription mapping into one segment of the hand transcription. Deletions were defined as two or more adjacent segments of the hand transcription mapping into one segment of the automatic transcription. (See Appendix C for insertion and deletion errors.)

Substitutions were defined to occur any time the hand and automatic labels did not agree, based on the 50% overlap criterion, independent of whether or not an insertion or deletion occurred. This definition was used because the match statistics used by lexical access were computed this way. The confusion matrices of Figure 3.8 show the substitution errors and correct labels. Some errors were due to the limited set of labels used in hand transcription. For example, the stop gap for /k/ in the word "six" was always transcribed as [k$^{\neg}$] if a demarcation was observed between the vowel and /s/, regardless of how noisy the closure was. Substituting "weak-fricative" for "silence" in this case is a reasonable error. Note that the bulk of the errors are reasonable, such as labeling /f/ (a weak fricative) as silence. Although /θ/ is also a weak fricative, in the digit lexicon /θ/ is always followed by an /r/. Since /r/ usually strengthens a preceding fricative, substituting strong fricative for weak fricative when a weak fricative is followed by an /r/ is also a reasonable error. A number of /n/'s were identified as a vowel. However, prevocalic /n/'s which may be a couple pitch periods in duration and therefore are not as salient as intervocalic /n/'s, are included in the statistics. Comparison of the performance for combinations of utterances and speakers reveals the performance to be similar. In cases where the labels differ, there are usually few samples, since these phones do not normally occur in digits. For example, voiced /h/ was sometimes used to mark aspiration at the end of sentence.

(a)

	Silence	Sonorant	Strong Fric	SVO	Vowel	Weak Fric
t°	96	2		0	0	1
k°	93		1		1	5
d°						100
ʊ	77	8	1	1	1	13
[]	32	11			4	54
w	3	75		3	19	1
n	8	70		4	16	1
s	1	0	90			8
z	1	7	78	1	2	12
t	9	1	56		1	34
k	41		52			7
v	15	23		31	21	10
ɾ		20		40	40	
ɛ		10		60	20	10
ə˞	0	0		0	92	
ɔ˞		13		3	84	
ü		14	1	2	81	1
ɪ˞		14		1	84	
ɐ˞		10		4	86	
ɔ	1	27		1	72	
u	1	20		2	68	
ʌ	1	1			99	
ɐ		20			80	
ɪ		2	0	0	97	
i					100	
ɛ					100	
r	1	6		0	92	0
ŋ	25	25			50	
ɱ					100	
h	9	39			17	35
ɦ		29			43	29
f	16	3	1		1	79
θ	13	5	12		0	69
+	33	28	6		11	22
ʔ	11	24		0	51	6
-		50			50	

(b)

	Silence	Sonorant	Strong Fric	SVO	Vowel	Weak Fric
t°	90	4	1		2	4
k°	91	1	4		1	4
ʊ	72	3			7	17
[]	100					
w	2	56		4	37	1
n	9	53	0	6	28	4
s	1		92		2	5
z	1	1	88		1	10
t	4	1	70		3	23
k	46	1	40		5	7
v	11	24	0	18	35	11
ə˞	1	1			98	
ɔ˞	1	9		2	89	
ü	3	20		4	73	
ɪ˞	3	14			83	
ɐ˞	1	10	2		87	
ɔ		27		1	72	
ʌ		3			97	
ɐ		6		1	93	
ɪ		2			98	
ɛ					100	
ɜ˞					100	
r	1	3		0	96	
ŋ		33			60	7
h	19	32			37	12
ɦ		40			40	20
f	20	4	2		1	73
θ	22	5	16		1	56
-	50		50			
ʔ	21	14	2	2	55	5
+	40		7		20	33
	50				50	

(c)

	Silence	Sonorant	Strong Fric	SVO	Vowel	Weak Fric
t°	94				3	3
k°	85					15
d°						
ʊ	76				12	12
[]		100				
w		71			29	
n	6	49		6	39	
s	1		90			9
z			74		9	17
t	7		43		4	46
k	50		50			
v	26	21		15	34	4
ɛ					100	
ə˞		2			98	
ɔ˞		5			95	
ü	0	8			85	
ɪ˞		12		7	80	
ɐ˞		9			91	
ɔ		35			65	
u					100	
ʌ		4			96	
ɐ		9			91	
ɪ		3			97	
ɛ					100	
r	1	7		1	91	
ŋ		60			40	
h	14	21	4		14	46
f	28		3			69
θ	8		18		5	69
-	33	33			33	
+	33	33				33
d					100	
ʔ	8	27		5	51	8

(d)

	Silence	Sonorant	Strong Fric	SVO	Vowel	Weak Fric
t°	91	2	2		2	4
k°	83	1			1	15
ʊ	68	8	3		5	16
[]	100					
w	2	79		1	17	
n	6	58		3	31	1
s	3		90		1	7
z	7		81		2	11
t	9		54		3	34
k	43		49			8
v	26	22		17	32	2
ə˞	2	3		2	94	
ɔ˞	1	5		4	90	
ü	2	17		2	78	
ɪ˞	1	8			90	
ɐ˞		8		3	88	1
ɔ		46		1	53	
u		44			56	
ʌ		5			95	
ɐ	2	12			86	
ɪ		3		0	95	1
ɛ					100	
ɜ˞				33	67	
r	0	9		1	89	
ŋ		57			43	
h	25	20			28	26
ɦ					100	
f	25	6	3		2	64
θ	22	6	15		3	54
+	25	38	25			13
ʔ	5	18	1	9	63	4

In addition to IPA symbols, additional symbols were used in the transcriptions: [] for silence within a word; ʊ for silence between words; + for noise; and - for voicebar.

Figure 3.8: Broad Phonetic Labeling Confusion Matrices for: (a) training utterance by training speakers (b) new utterances by training speakers (c) training utterances by new speakers (d) new utterances by new speakers

3.2 Lexical Access

Based upon the continuous speech recognition model, the lexical access component produces viable word hypotheses using its knowledge of the words in the lexicon, path constraints, how allophonic realizations of a phone are context-dependent, and reasonable durations for speech units. The component was implemented in two parts. First the word hypothesizer used sequential constraints to propose word candidates. Second, the word candidate pruner used durational, allophonic and path constraints to reduce the number of viable word candidates. The lexical access component could have been implemented so that checks on some of the pruning conditions are made as words are hypothesized. But by performing hypothesizing and pruning separately, the source of any errors may be more easily found. Since the purpose of developing the component was to study how speech knowledge could be applied to real speech for continuous speech recognition, the ability to quickly understand the source of error was important; consequently each constraint is applied as a separate step.

Lexical access produces a lattice of word candidates for the verifier. The application of speech constraints to prune the lattice produced by the word hypothesizer is important because the verifier can focus on reasonable candidates without performing computations on unreasonable word hypotheses, thus resulting in a more directed search.

3.2.1 Dictionary Representation

A word may be pronounced in multiple ways. The lexicon contains knowledge about allowable pronunciations of each word and the context in which each word can occur. An average of two pronunciations per word was used. Sentence-initial pronunciations allowing for a voice bar in "zero", a very weak initial /f/ in "five", and no initial closure in "two," were used. In addition, sentence-final pronunciations

allowing for aspiration following the final /r/ in "four" and for deletion of the final closure in "eight" were used. Each phonetic pronunciation of a word is stored in an association list which is keyed by phonetic transcription. Associated with each pronunciation is a structure containing broad contextual information. This information is divided into four parts:

1. broad classes which must follow or cannot follow the current pronunciation. For example, [e^{y}ɾ] requires a following vowel.

2. broad classes which must precede or cannot precede the current pronunciation. This information was not used because it was not applicable to the pronunciations used.

3. whether the first phone can geminate. The first phone in the pronunciation is not allowed to geminate when the first phone in the canonical pronunciation is deleted. For example, when the nasal murmur in the initial /n/ in "nine" is deleted, the /a^{y}/ is not allowed to geminate with a preceding vowel, since intervocalic /n/'s are usually robust and should be found by the classifier.

4. whether the last phone can geminate. For example, a flap, as in [e^{y}ɾ] is not allowed to geminate.

When a word is hypothesized, this information is kept with the word and accessed at the appropriate point in processing.

3.2.2 Hypothesizing Words

The word hypothesizer produces viable word candidates given the broad phonetic segmentation produced by the broad phonetic classifier and knowledge about the words which form the lexicon. In the ideal case where interspeaker and intraspeaker variations are minimal and the broad class segmentation is accurate, sequential constraints can be applied directly to the segmentation string. That is,

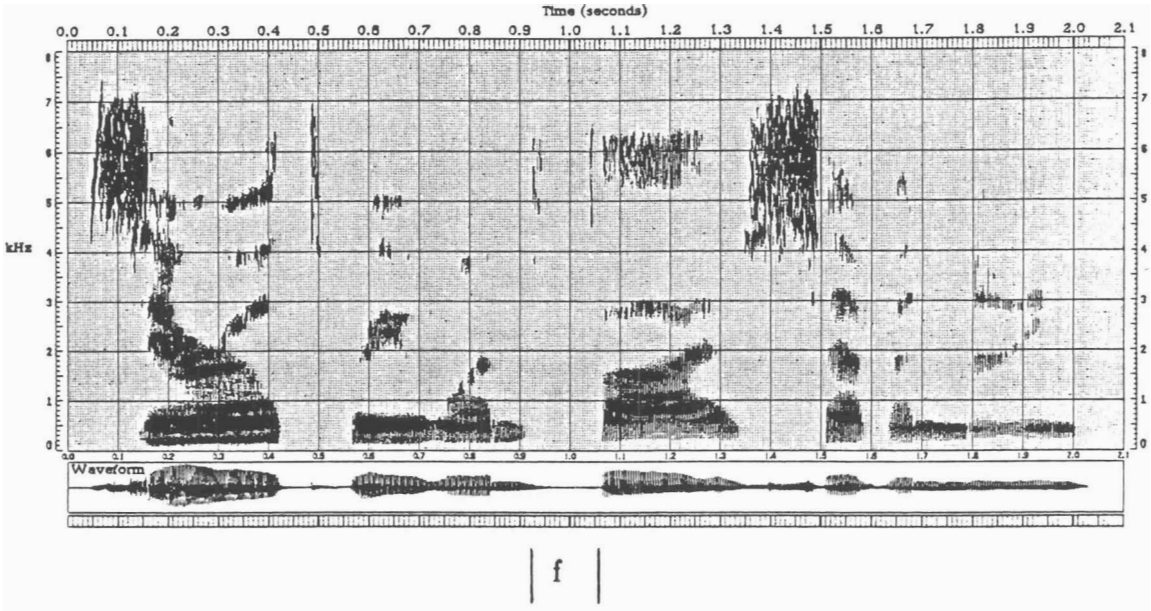


Figure 3.9: Spectrogram of the digit string "031579" with silence in /f/ in "five"

for each pronunciation of a word in the lexicon, which is represented by its phonetic and corresponding broad phonetic transcriptions, matches are found between a portion of the segmentation string and the broad phonetic representation of the pronunciation of a word.

Use of a segment lattice in place of the segmentation string can reduce the possibility of incorrect segmentation of real speech; however, a lattice handles speaker variations by enumeration. Different pronunciations produced by different speakers, must be included as alternate pronunciations, resulting in a very large lexicon. Furthermore, given an unanticipated realization of a word, the system will be unable to correctly propose the word. For example, a short period of silence in the middle of the /f/ in "five" may have been observed (see Figure 3.9), but the /f/ in "four" may have never been observed to be pronounced this way. If a speaker then says the /f/ in "four" with a short period of silence in the middle, the system should use the knowledge that it has seen /f/'s in other words pronounced this way. Thus the system should give the /f/ a good score, rather than commit a fatal error by

assuming that the /f/ in "four" could never have some silence in the middle.

The use of a lattice was examined by performing lexical access on the segment lattice produced by the broad phonetic classifier immediately before the lattice is reduced to a segmentation string. The performance was evaluated on five new speakers by measuring how often the correct word was not among the word candidates. A word is considered a candidate when it overlaps in time with the spoken word by more than 50%. The correct digit was not one of the lexical candidates only 1% of the time. However, this measure did not insure that path constraints would be satisfied. For example, /θri/ is represented as "weak-fricative vowel" at the broad phonetic level. If the underlying /θ/ is classified as "weak-fricative silence weak-fricative" by the broad phonetic classifier, and this representation was not in the lexicon, then "three" would be a lexical candidate, but the "weak-fricative silence" portion of the segment lattice would not be associated with the three. Thus an alternate pronunciation of /θ□θri/ would have to be added to the lexicon. It was found that a segment lattice did not provide enough flexibility and that the size of the lexicon had to be increased to accommodate new pronunciations.

Thus the word hypothesizer used knowledge about the characteristics of the segmentation strings produced by the front end. A scoring algorithm was developed to allow for some acoustic variations in a phone. Many alternate pronunciations needed with the straight matching method are unnecessary with this method because the system knows the types of errors that the broad phonetic classifier tends to make and uses that knowledge in scoring each word. For instance, the broad phonetic classifier may call /θ/ a weak fricative 60% of the time and a strong fricative 40% of the time. The system knows this and therefore when a /θ/ is called a strong fricative, the score is not reduced much. In contrast, if a /θ/ is never called a vowel, then the match of /θ/ to the broad class label "vowel" would be assigned a poor score. In addition to substitution errors, the algorithm also handles insertion and deletion errors by using transition probabilities. If the broad phonetic classifier

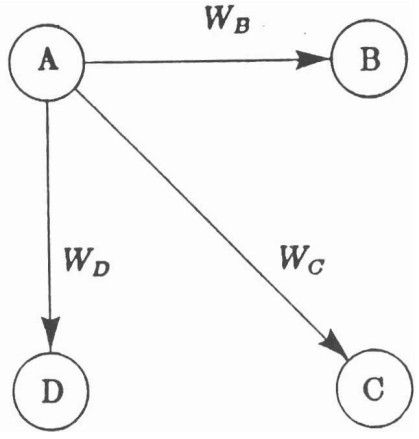


Figure 3.10: Paths Used in Dynamic Programming Algorithm

consistently misses prevocalic nasals, as in the word "nine," then the system will know that very often the /n/ as well as the /aʸ/ is labeled as a vowel. This is reflected by high transition probability of matching /n/ to "vowel" and then matching /aʸ/ to "vowel."

The general idea of the matching algorithm is to use knowledge about the characteristics of the broad phonetic classifier's output to assign a score reflecting how well a phonetic pronunciation matches a portion of the segmentation string. Each segment of the segmentation string is used as a beginning segment for matching each pronunciation. One or more end segments are associated with a beginning segment. Each end segment is chosen based on the length of the pronunciation's phonetic transcription; the end segments are iteratively chosen to be all segments within the range: $S_b + fixr(.5 * L)$ and $S_b + fixr(2 * L)$ where S_b is the beginning segment, L is the number of elements in the phonetic transcription, and $fixr$ is the operation of rounding to the nearest whole number.

A forward dynamic programming algorithm (Winston, 1984) was used to find the best match between the two sequences. The allowed paths from each node are illustrated in Figure 3.10. Simple 1:1 slope constraints are used, requiring the path to be monotonically non-decreasing in each direction. In contrast to the constraints

used in dynamic time warping of the speech signal, many phonetic segments may map into a single label, as the /ɪ/, /r/ and /oʷ/ in "zero" maps into the label "vowel," because the broad phonetic classifier has no knowledge for differentiating among these sounds. Figure 3.10 shows that three paths or transitions (to nodes B, C, and D) exit from a typical node A. The total accumulated score or "distance" to node D, d_D is computed as:

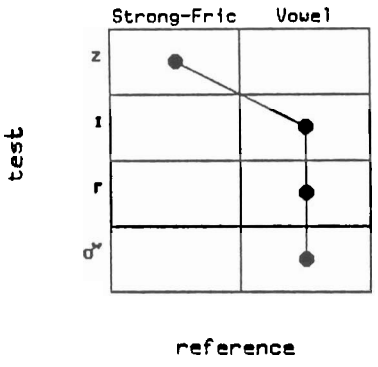
$$d_D = d_A + log[\Pr(p_D, l_D) * W_D]$$

d_A is the total accumulated score to node A, $\Pr(p_D, l_D)$ represents the probability of the phone at node D, p_D, being labeled as the broad class label l_D. W_D is the probability of making a transition from node A to D, given that node A is the current state and nodes B,C, and D are states which may be entered from node A. W_D is computed as:

$$W_D = \frac{\Pr(p_A l_A \rightarrow p_D l_D)}{\Pr(p_A l_A \rightarrow p_B l_B) + \Pr(p_A l_A \rightarrow p_C l_C) + \Pr(p_A l_A \rightarrow p_D l_D)}$$

Thus W_D represents the probability of deleting a segment such that p_A and p_D map to l_A, which is the same as l_D. Similarly, W_B represents the probability of deleting a segment. The use of these weighting functions incorporates information about insertion and deletion probabilities into the score.

Figure 3.11 illustrates the alignment between the phonetic string /zɪroʷ/ and the broad phonetic representation "strong-fricative vowel." Probability scores used in the computation are shown on the right. The alignment score between each phone and broad phonetic class is shown under "match." The transition score from the previous node to the current node is shown under "weight." A weight is not shown for the first phone and broad class pair because a transition was not made. Not that an insertion or deletion occurs 1% of the time in the first transition only. The total accumulated score to a phone and label pair is shown under "total." The score assigned to a phonetic string is the total score of the best path. This score is normalized by the number of transitions and is shown as the final score in the figure.



reference	test	match	weight	total
STRONG-FRIC	z:	0.7754	1.0000	0.7754
VOWEL	I:	0.9714	0.9900	0.7457
VOWEL	r:	0.9236	1.0000	0.6888
VOWEL	o:	0.8414	1.0000	0.5795
				0.8725

Figure 3.11: Alignment of /zɪro^{w}/ with "strong-fricative vowel"

By allowing each pronunciation of a word to begin at each segment with possible multiple end segments, many words can occur in a word lattice. For example, in a lattice represented by 20 broad phonetic segments, a word represented by four phones would be hypothesized 112 times. Although the number of candidates is large, this number is much smaller than the possible number of candidates in a frame-by-frame approach, such as dynamic time warping, where each word can begin at every frame. Furthermore, many of these hypotheses are poor matches, and some type of pruning can be applied to remove these poor matches.

Two word score distributions, on a $\log_e$ scale, are shown in each part of Figure 3.12. The distribution of correct word scores is indicated by the dashed line, and the distribution of the scores of all word hypotheses for a sample utterance are indicated by a solid line. Comparing the two distributions in each figure, we note that the log probability scores of the correct words are much closer to 0, or a probability of 1, than the bulk of the scores of all possible words. Note also that the distributions are similar for the new utterances spoken by training speakers and by new speakers, indicating the potential speaker independence of the approach.

A word score threshold can then be set such that all words with a score below the threshold are ruled out as a viable candidate. If a word is pruned as soon as

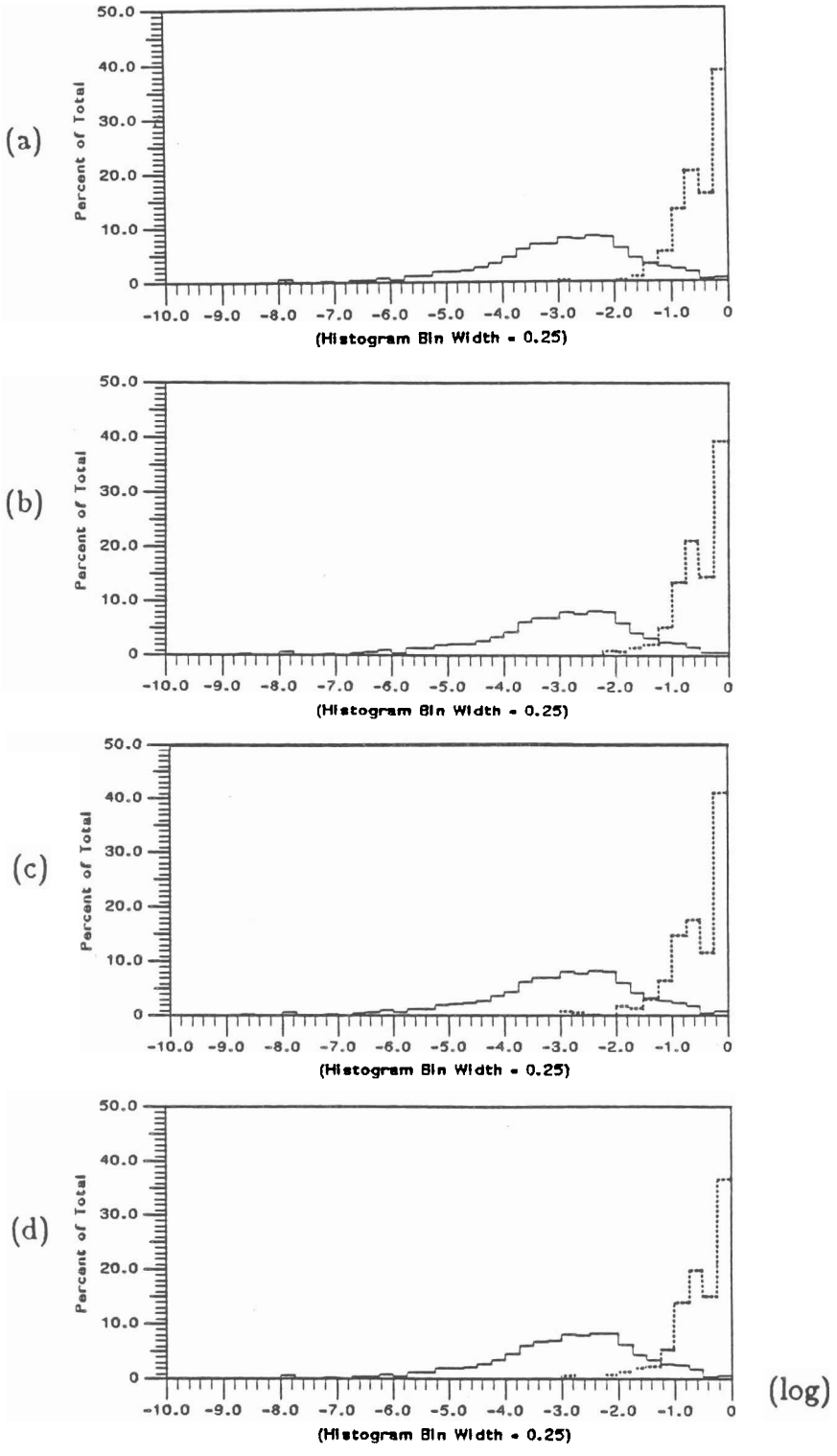


Figure 3.12: Histograms of Correct Word Scores vs All Word Candidate Scores: (a) training utterances by training speakers (b) new utterances by training speakers (c) training utterances by new speakers (d) new utterances by new speakers

the cumulative cost computed by the alignment algorithm passes the threshold, the amount of computation required for finding word candidates can be significantly reduced (by about an order of magnitude as measured from run times when the threshold is set at -1.5). Thus by setting the threshold at different values, the system can be biased towards less computation with more errors, or towards less errors and more computation, as desired.

Figure 3.13 illustrates the relationship between the amount of pruning achieved compared to the percentage of correct words pruned. We can observe the similarity of the curves and how the reduction of all word candidates is much more than the reduction of the correct word candidates for a given pruning threshold.

Figure 3.14 illustrates the relationship between the number of word candidates in the word lattice per word in the digit string as a function of the pruning threshold used in application of sequential constraints. Note that the number of word candidates increases sharply as the threshold is initially relaxed. The large number of word candidates when the threshold is very weak is due to the combinatorics of allowing words to begin and end at multiple segments, as explained earlier in this section.

3.2.3 Pruning the Word Lattice

Simple constraints based upon speech knowledge can be used to rule out very unlikely candidates. For example, if [eʸɾ] ("eight" pronounced with a flapped /t/) is hypothesized, then the following word must begin with a vowel, since a /t/ is flapped only in the context of vowels. If none of the following hypothesized words begins with a vowel, then [eʸɾ] can be ruled out as a viable word candidate.

Three types of constraints were applied following word hypothesis: path constraints, durational constraints, and allophonic constraints. The block diagram in Figure 3.15 illustrates when each constraint is applied. For example, durational constraints are applied first to rule out word candidates which depend on a seg-

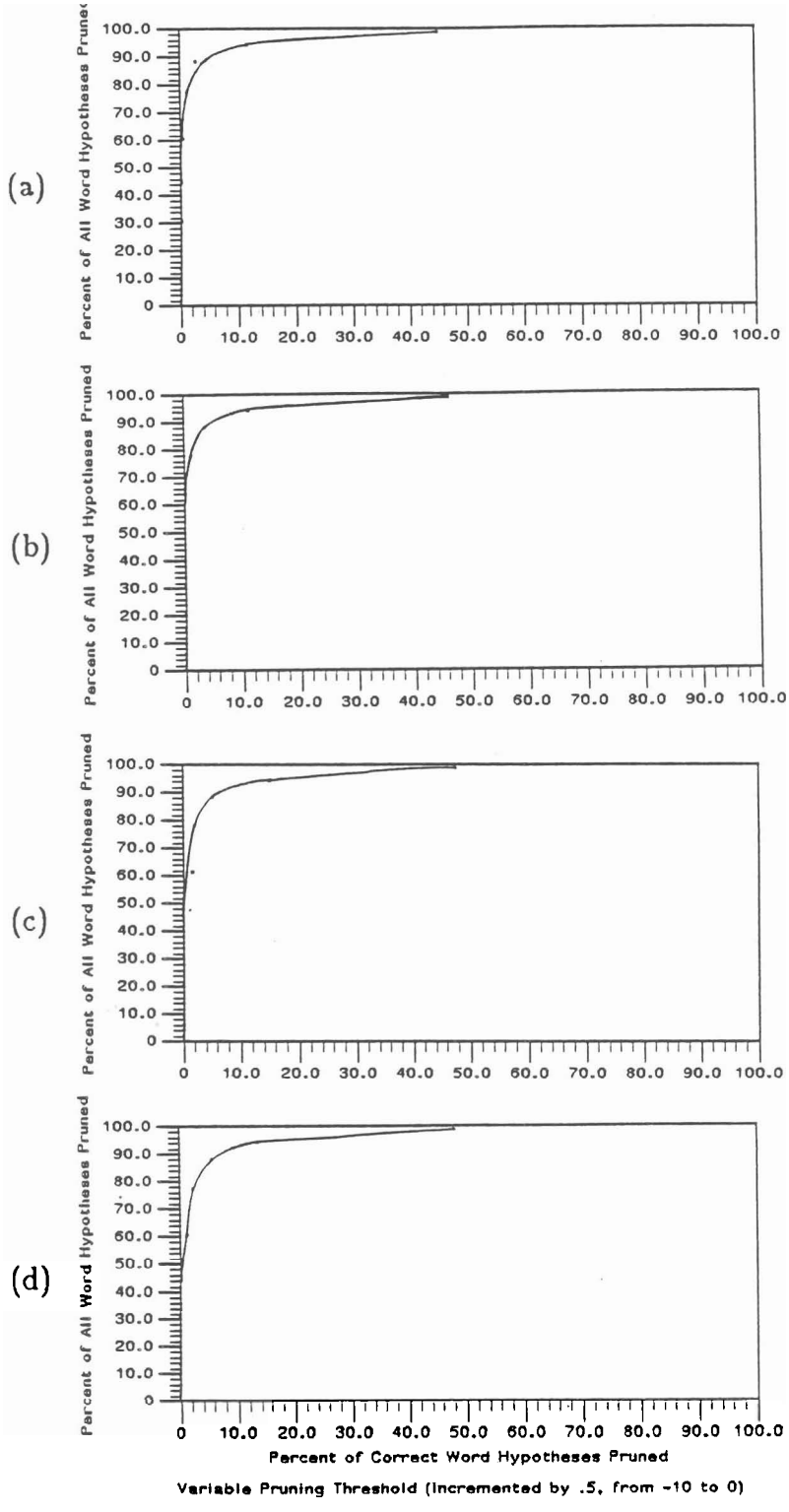


Figure 3.13: Pruning of All Word Hypotheses and Correct Word Hypotheses (a) training utterances by training speakers (b) new utterances by training speakers (c) training utterances by new speakers (d) new utterances by new speakers

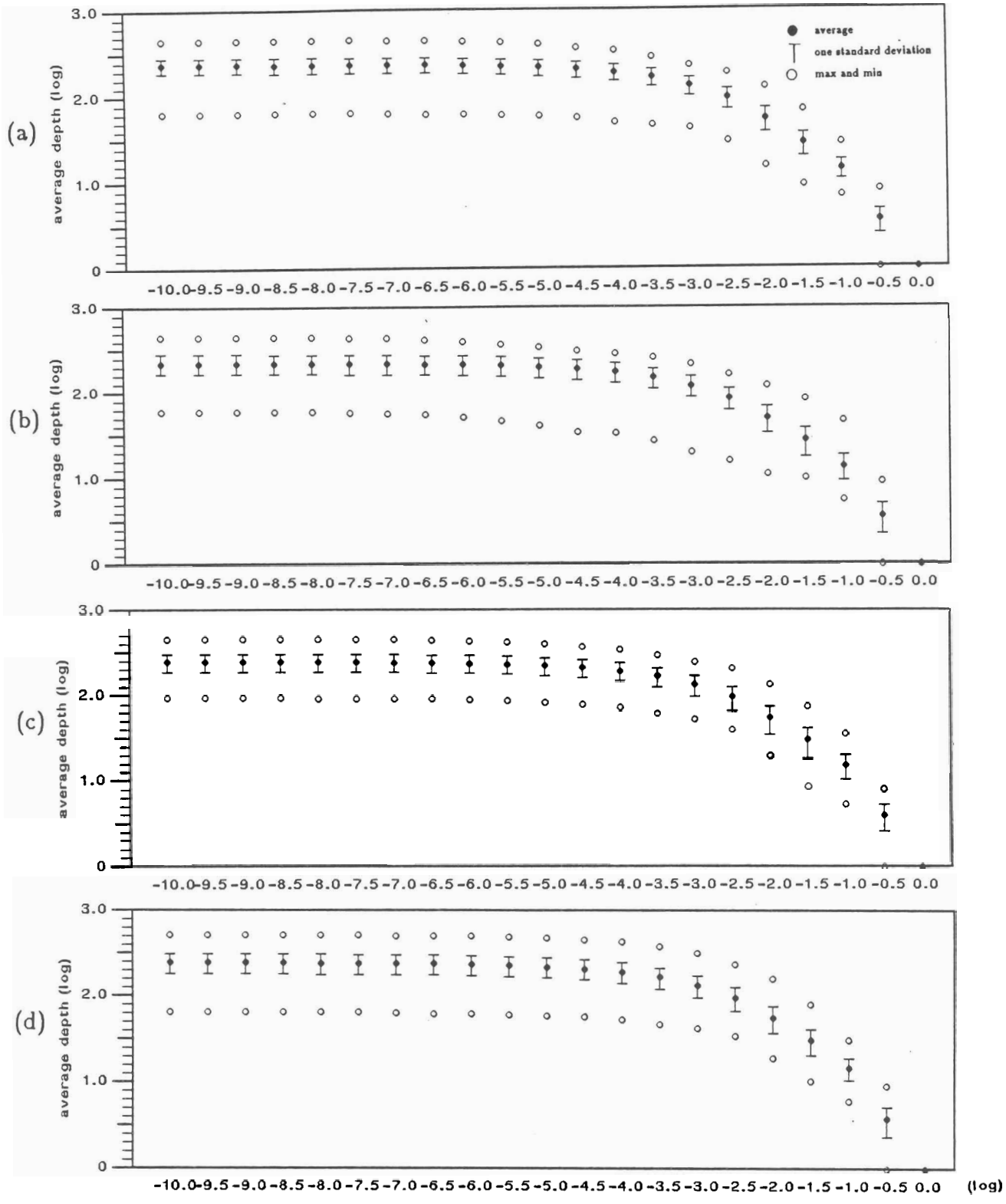


Figure 3.14: Lattice Depth vs Sequential Constraint Threshold (a) training utterances by training speakers (b) new utterances by training speakers (c) training utterances by new speakers (d) new utterances by new speakers

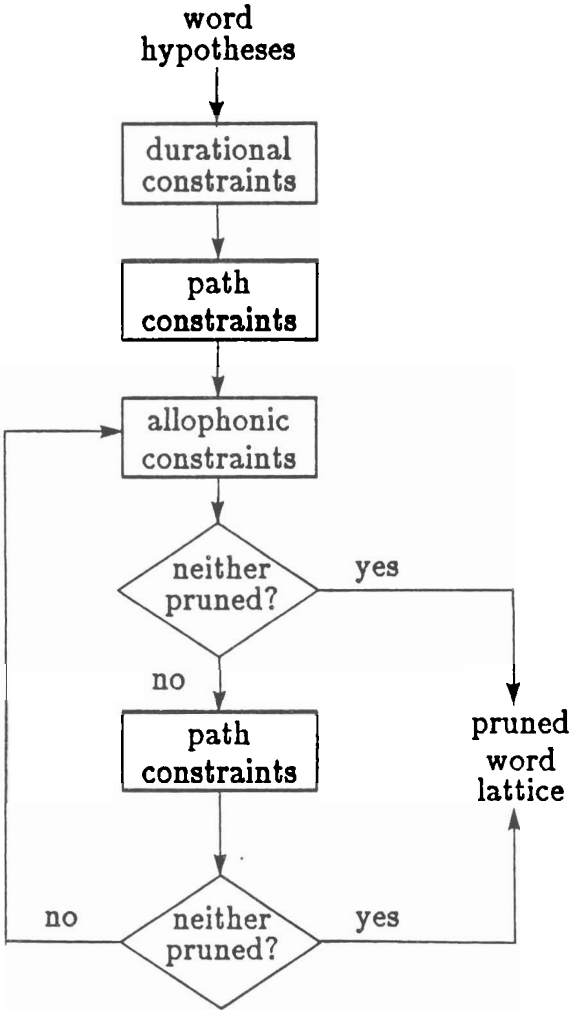


Figure 3.15: Application of Pruning Constraints

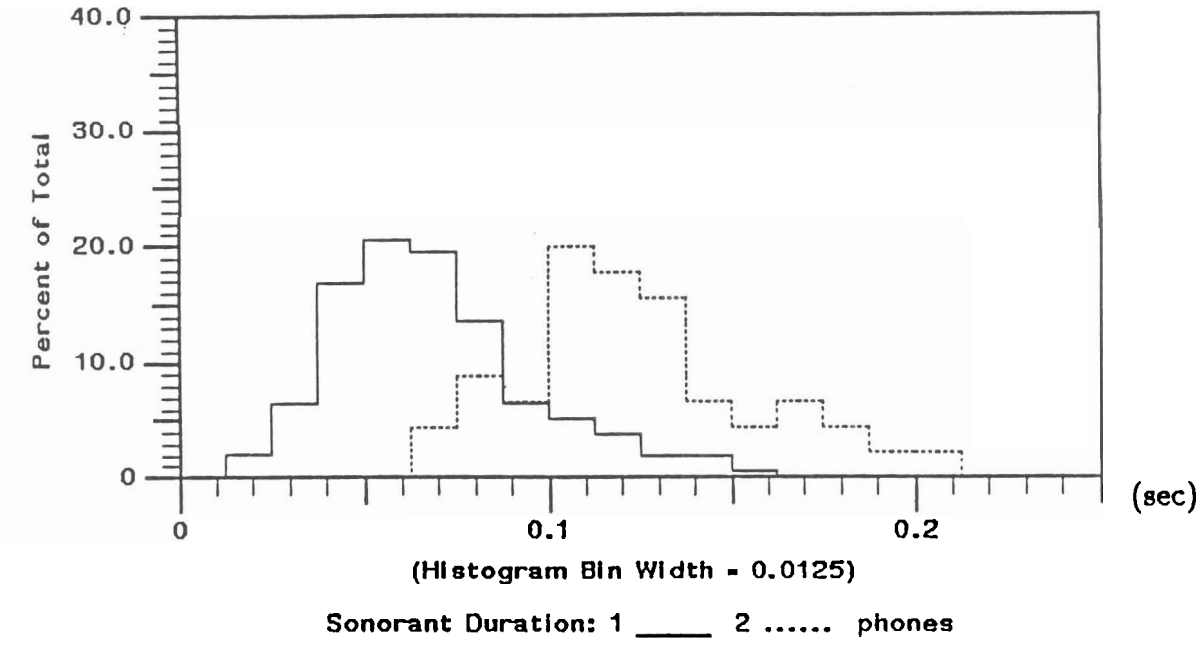


Figure 3.16: Distribution of sonorant duration for segments containing 1 or 2 phones

ment representing 1 phones when the durational training data indicates that 2 phones are represented by the segment. After duration constraints are applied, path constraints and broad allophonic constraints are applied iteratively until neither constraint prunes any word from the word lattice. Path and broad allophonic constraints are applied iteratively because removal of a word by one constraint may cause the conditions for the other constraint to remove a word(s) to be satisfied.

As an example, Figure 3.16 shows the distribution of sonorant segment durations in the training set (see Appendix A for a description of the training set). Table 3.4 shows the cutoff points for regions where only 1 phone, 1 or 2 phones, and only 2 phones occurred for the broad phonetic classes: "sonorant," "vowel" and "strong fricative."

Path constraints, as described in Section 2.3.2, require that each word in the lattice form part of a complete path. Words which do not have a legal "next word" and "previous word" are pruned from the lattice. The "next word" can be either a word which begins where the current word ends, the end of the sentence, or a word that has an initial phone which could acoustically geminate with the final phone of

Table 3.4: Cutoff Points of Segment Duration

Broad Class	Lower Cutoff (sec)	Upper Cutoff (sec)
Sonorant	.04	.17
Strong Fricative	.09	.15
Vowel	.09	.26

the current word; "previous word" is defined similarly. Acoustic gemination is the merging of two phones such that they are acoustically realized as one phone. This occurs when two adjacent phones are similar, such as the final /s/ in "six" and the initial /z/ in "zero." These two phones may merge such that they appear as one strong fricative.

Broad allophonic constraints, as described in Section 2.3.4, require that the context of a word be compatible at the broad phonetic level. The flapped /t/ in "eight" is an example of broad allophonic constraints. When none of the word hypotheses satisfy the contextual constraints of a word, the word can be removed as a viable word hypothesis.

Figure 3.17 illustrates the primary steps in lexical access. Each "box" in the figure represents the relative position of a label and does not convey any information about duration or rank order. The broad segmentation is shown in (a). Below in (b), all word candidates with scores better than the pruning threshold (chosen to be -0.75 for this example) are displayed in the word lattice. Application of duration constraints did not remove any word candidates in this example. The first application of path constraints removed the words marked with a sharp sign in (b), producing the word lattice shown in (c). All the words ending at the last vowel segment which could not geminate with a word beginning in the vowel segment, denoted by a "#" in (b) were removed because no word in the vocabulary is composed only of a sonorant. The first application of broad allophonic constraints removed

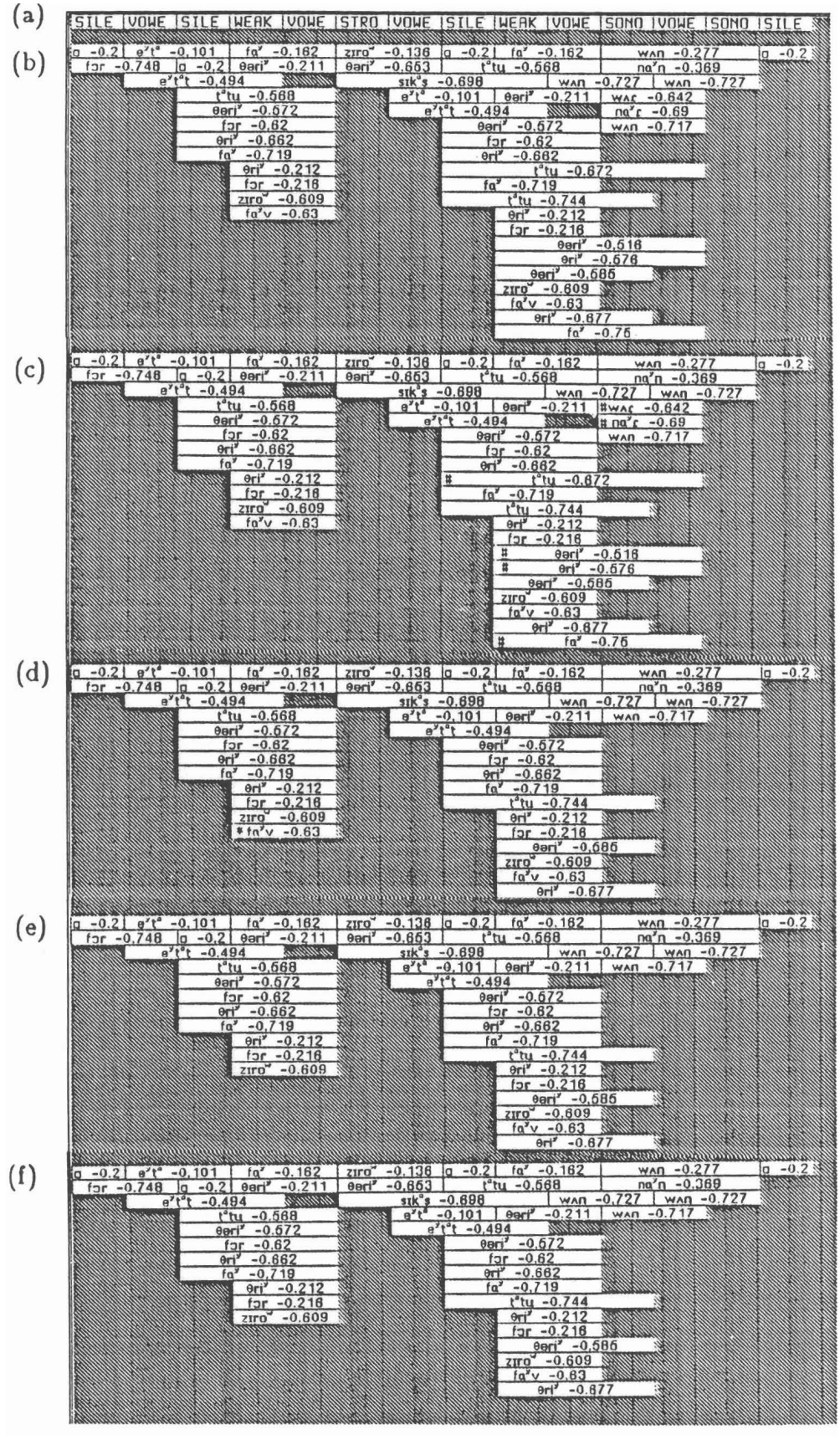


Figure 3.17: Pruning Word Candidates in Lexical Access

the /faʸv/ marked with an asterisk in (d), producing the word lattice shown in (e). The /faʸv/ was removed because in the context where a strong fricative follows /v/, the /v/ may be realized only as a vowel or fricative. Thus the pronunciation of "five" with the /v/ was removed, but the pronunciation with the /v/ deleted was kept. The final pruned lattice is shown in (f).

The pruning constraints were evaluated on the utterances in the training set by comparing the number of word candidates immediately after all words are hypothesized with the number of word candidates after pruning. Two sequential constraint thresholds were tested. The ratio of word candidates after pruning to word candidates before pruning was 0.69 at a sequential constraint threshold of -0.75 and the ratio was 0.76 at a sequential constraint threshold of -1.75. Thus application of durational, path, and allophonic constraints can reduce the number of word candidates even further. However, the constraint is not as strong when the sequential threshold is more lenient. With a weaker threshold (more negative), additional word candidates are allowed which can prevent the pruning constraints from being satisfied. Therefore, the pruning constraints need some prior reduction of word candidates before they can be applied effectively.

3.3 Chapter Summary

The main points addressed in this chapter are:

- Speech is highly variable. However, this variability is less evident in a broad phonetic representation. Thus many of the variabilities can be handled by representing speech at a broad phonetic level.

- The broad phonetic classifier segments and labels speech into broad phonetic classes using a set of production rules applied to coarse acoustic features. The broad acoustic features characterizing the speech signal are defined by identifying robust regions and then extending outward.

- A delayed binding approach is used to produce the final segmentation string; that is, the segment string is determined after all candidate labels are found.
- Even with a segment lattice, the lexicon becomes very large due to speech variations
- By using the characteristics of the broad phonetic classifier, sequential constraints can be effectively applied to real speech.
- Path constraints and broad allophonic constraints can be applied to reduce the number of candidates even further.

Chapter 4

Feature-Based Verification of Word Hypotheses

This chapter explores the use of detailed acoustic features for verification of word hypotheses. In our model, the input to the verifier is a lattice of word candidates produced by the lexical component. In this lattice, the most unlikely candidates have been removed. The task of the verifier is to select the best word or string of words from among the competing word candidates using a set of detailed acoustic features.

To verify the word candidates, each word hypothesis is represented as a sequence of phones, and each phone is characterized by a set of detailed acoustic features. The features were chosen to capture salient acoustic characteristics of speech sounds and to provide good discrimination among easily confused sounds in the digits. Two simple scoring algorithms were then used to demonstrate the feasibility of using these acoustic features for verification.

4.1 Characterization of Phones

Although many different recognition units could be used to score word hypothe-

ses, phones were chosen as the basic recognition unit because they are suitable for defining many types of phonetic features and because of potential extendability to other recognition tasks. By representing each of the word hypotheses as a sequence of phones, features can be developed to characterize phones rather than whole words. Since the number of phones in a language is limited, a phone representation is potentially extendable.

Many of the defined features take advantage of the fact that a phone representation is used. A phone representation allows features to be defined over specific regions of time. In the middle of a phone region, characteristics of only one type of sound are present, while at the edges of the region, transitional information is available. Thus features can be defined to be minimally influenced by coarticulation or to specifically capture coarticulation effects. For example, the features for characterizing F_1 and F_2 were defined such that coarticulation effects on the computed values are minimal.

A phone representation also allows a wider variety of acoustic-phonetic constraints to be exploited more easily than a frame-by-frame representation. A phone can be characterized over its entire duration, such as by the maximum or average values of a set of parameters, rather than checking values on a frame-by-frame basis. In addition, many characteristics of speech sounds which are difficult to capture in a frame-by-frame approach are easily captured in a phone representation. For example, onset rate characterizes a phonetic event, such as a rapid stop release. This acoustic event may occur in one or two frames in a frame-by-frame approach and influences the overall score only in the one or two frames. In a phone representation, such speech events can be captured explicitly, and in an acoustic-phonetic approach, the information can be given equal weight with other feature information. Thus the onset rate can be given more import in the decision process in a phone representation than in a frame-by-frame approach.

A small set of acoustic features was carefully designed to capture salient char-

acteristics of phones and detailed differences between similar phones. The large amount of information needed to represent a spectrogram can be reduced by defining and identifying acoustic features which robustly capture the occurrence of significant events in the spectrogram. Observations of phone characteristics in a spectrogram were used to select the initial parameters and features. The parameters were chosen to characterize important regions or events in the speech signal, and the features were chosen to quantify the parameters within a phone region.

4.1.1 Parameters

Three types of parameters, the energy within a frequency band, a measure of the location of spectral concentration, and the location of the offset of the first major peak in smoothed spectra, were used to characterize the speech signal. All parameters, except for the peak offset, were computed at a fixed rate. The information each parameter captures and how each parameter is computed are described below.

Energy

The energy in a frequency band, E, is computed by applying a frequency window, $\vec{W}(e^{j\omega})$, to the short-time spectra, $\vec{X}(e^{j\omega})$:

$$E = \log\left[\vec{X}(e^{j\omega}) \bullet \vec{W}(e^{j\omega})\right]$$

Trapezoidal shaped frequency windows were used to minimize sharp changes in energy as a formant moved in or out of the frequency band. The energy was computed as log energy to minimize the sensitivity to small variations in value when the value is large.

A Hamming window, h[n], was used to calculate the short-time spectra:

$$\vec{X}(e^{j\omega}) = \sum_{k=-\infty}^{\infty} x[k]h[n-k]e^{-j\omega k}$$

Unless stated otherwise, the duration of the Hamming window was 25.6 msec. By varying the width of the Hamming window and the default update rate of 5 msec, the sensitivity of the parameter to energy changes in the speech signal can be modified.

Spectral Concentration

One of the primary characteristics of voiced phones is the presence of formants; and associated with each formant is a concentration of energy. When two formants are close in frequency, these energy concentrations may merge so that accurate tracking of formant frequencies is difficult. However, in the identification of phones, particularly the limited number of phones in the digit vocabulary, gross characteristics of energy location may be sufficient.

Spectral weighting windows were used to characterize the location of energy concentrations associated with the formants of speech. The location of spectral concentration, S, was computed by applying a weighting window to the spectrum:

$$S = \frac{\vec{X} \bullet \vec{W}}{\sum_i |x_i w_i|} = \frac{\sum_i x_i w_i}{\sum_i |x_i w_i|} \tag{4.1}$$

where the magnitude-squared spectrum, $\vec{X}$, and weighting window, $\vec{W}$, represent vectors normalized by the mean value of the vector elements. The vectors are normalized by the mean value of the vector elements to remove bias in the computed value. Additional normalization by the ***magnitude*** of each pair of vector elements was motivated by the idea that when x_i and w_i are similar, they will add constructively as $|x_i w_i|$, but when x_i and w_i are different, they will add destructively relative to $|x_i w_i|$. Thus if $\vec{X}$ and $\vec{W}$ are very similar, S will be close to 1; if $\vec{X}$ and $\vec{W}$ are very dissimilar, S will be close to -1.

The weighting window can be any real function over the frequency range; it may be thought of as a generalized form of the center of mass or first moment, in which the weighting function is linear with frequency. By specifying the sensitivity of the window to different regions of the spectrum, the weighting function can be tailored

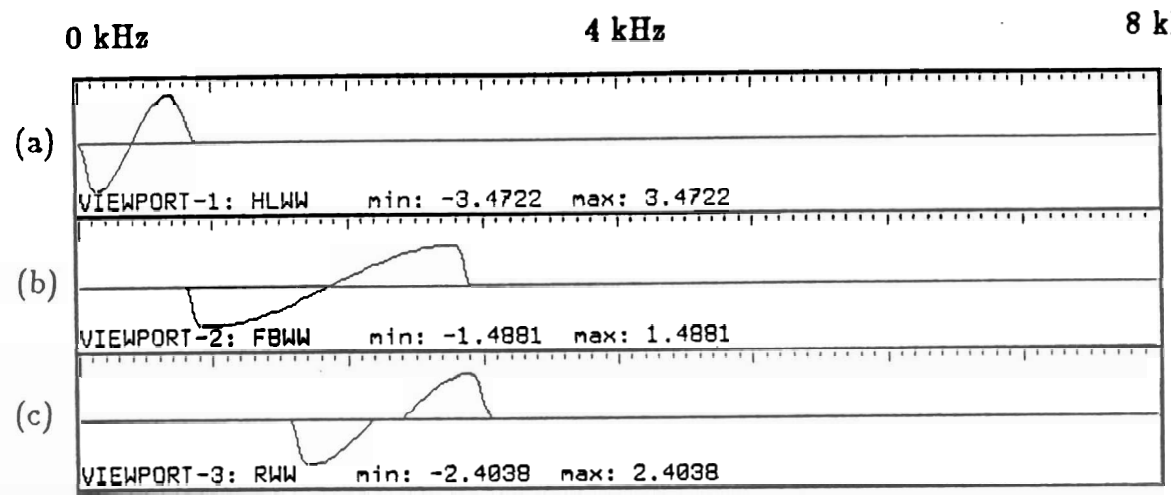


Figure 4.1: Weighting Windows: (a) High/Low (b) Front/Back (c) /r/

to capture particular spectral features. The weighting windows were tailored to be least sensitive near the edges of the frequency region in order to minimize the effect of formant motion at the edge of the windowed region. This was accomplished by defining weighting windows with a slope of zero at the outer edges.

Figure 4.1 illustrates the three weighting windows used. These window positively weight the higher frequencies and negatively weight the lower frequencies. Thus the lower in frequency the energy is concentrated within the band, the lower the value of the spectral concentration. The first weighting window captures the position of energy concentration in the frequency range below 900 Hz. In vowels, this "high/low-spectral-concentration" parameter[1] roughly corresponds to the position of the first formant; and in nasal consonants, the value of this parameter is a function of the position of the nasal formant, nasal zero, and the first formant. The second weighting window provides information on the location of energy concentration in the frequency range from 900 Hz to 2800 Hz. This "front/back-spectral-concentration" parameter roughly corresponds to the position of the second formant in vowels and glides. The third weighting window attempts to capture when the

[1]Some of the names given the parameters, features, and terms used in this chapter were chosen for conciseness at the sacrifice of accuracy. For example, the high/low-spectral-concentration does not always correspond with whether or not the vowel is high since this parameter may be influenced by F_2 when it is low in frequency.

third formant lowers in frequency for an /r/. During the production of an /r/, the third formant characteristically drops to approximately 2000 Hz. The exact frequency depends upon a number of factors, including the speaker and speaking rate. The amount of lowering also depends on whether the /r/ is prevocalic, postvocalic, or intervocalic; all three types of /r/ occur in the digit vocabulary. The "/r/-spectral-concentration" parameter looks for the presence of energy around 2000 Hz. Since the exact frequency to which F_3 dips is speaker-dependent, a "null" region was designed into the window, giving less import to exactly how low F_3 dips.

The high/low-, front/back-, and /r/-spectral-concentration parameters were designed to provide information about F_1 and F_2 and information about when F_3 lowers for an /r/. Sample output for the three weighting windows are shown in Figure 4.2. The /r/-spectral-concentration parameter approaches -1 when an /r/ is present. The values of the high/low- and front/back-spectral-concentration parameter usually correspond with F_1 and F_2, although the parameter values appear to change sharply. This is due to the shape of the window, which is more sensitive in the transition regions. In addition, note that the value of the high/low-spectral-concentration parameter at the end of the /w/ is larger than expected because of influence by F_2.

Offset of First Peak in Smoothed Spectra

The location of the offset of the first major peak in the shape of the spectrum was computed by finding the upper edge of the first peak in a cepstrally smoothed spectrum. Cepstrally smoothed spectra (Oppenheim and Schafer, 1968) were computed by applying a quefrency lifter which is constant the first 0.7 msec and cosine tapered the next 1.0 msec to the cepstrum of a 1024-point DFT. The DFT spectra were computed for the frequency range from 0 to 8 kHz. The extreme smoothing of the low-pass window produces a spectrum in which pitch harmonics are removed and only gross characteristics are evident. This type of spectral representation

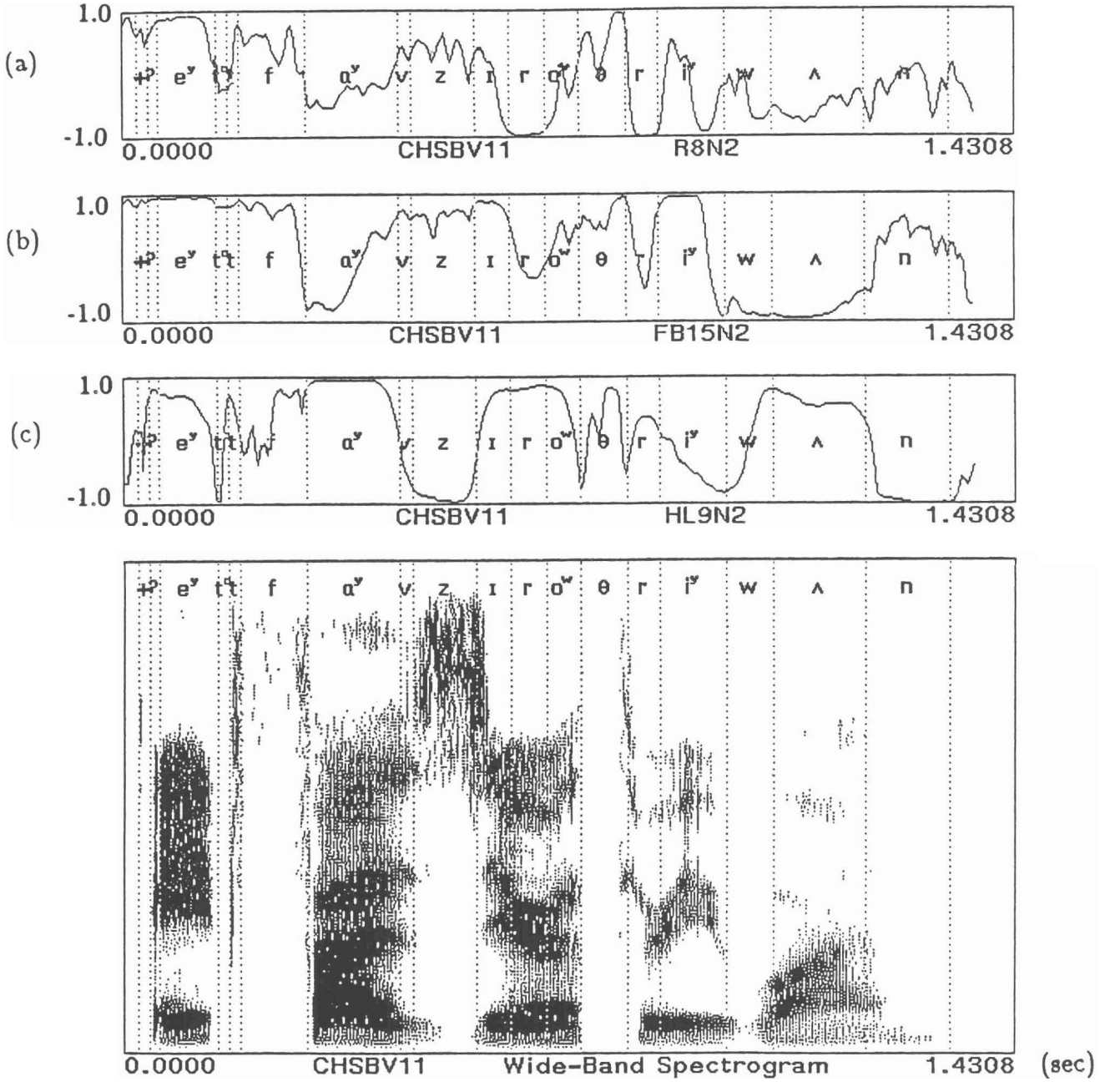


Figure 4.2: Sample Spectral Concentration Output for the Digit String "85031" (a) /r/-spectral-concentration (b) front/back-spectral-concentration (c) high/low-spectral-concentration

simplifies characterization of general spectral shape.

To locate the offset of the first major spectral peak, the smoothed spectrum was searched upward in frequency from the first peak until a value 12 dB down from the peak value was found. If a larger peak was encountered during the search, the peak value was set to the value of the larger peak. This parameter was computed for only one point per phone, partially because of the large amount of computation required to produce the smoothed spectrum. But unlike DFT spectra, this representation is relatively stable from sample to sample so that a high update rate appeared to be unnecessary.

4.1.2 Features

The values of a parameter within a phone region were converted to one feature value which characterizes some aspect of the parameter within the region. Nine acoustic features were defined for discriminating the digit phones. The nine features roughly describe the first three formants, the presence of a low frequency nasal pole, the onset rate of phones, the upper frequency of the first major concentration of energy, and changes in the noise source energy over the duration of a phone. A short name by which the feature will be referred to in later sections is given below for each feature. A more precise name and a description of the computation of each feature then follow. The nine features are:

1. **F_1-Normalized-Position:** This feature is the average value of the high/low-spectral-concentration over the middle 50% of the phone region:

$$\frac{\sum_{t=t_{.25}}^{t_{.75}} S_{F_1}(t)}{t_{.75} - t_{.25}}$$

where $S_{F_1}(t)$ is the high/low-spectral-concentration at time t, $t_{.25} = .25d_p + t_b$, $t_{.75} = .75d_p + t_b$, d_p is the duration of the phone in number of samples, and t_b is the sample at the time the phone begins. This feature is most useful for identifying vocalic phones and usually indicates whether a phone is more

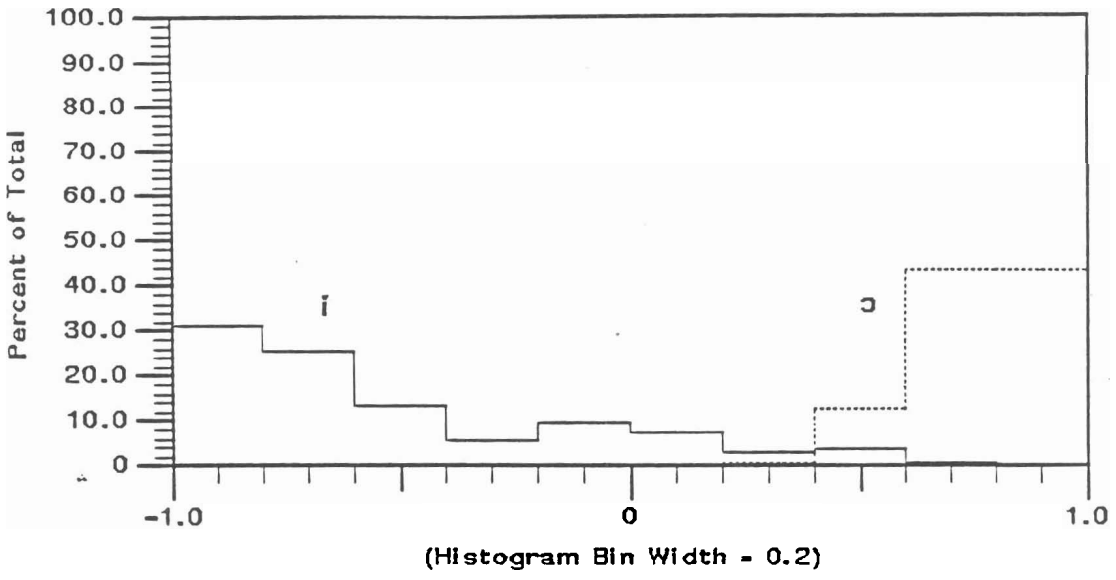


Figure 4.3: Distribution of F_1-Normalized-Position for /i/ and /ɔ/

like a high or low vowel. Figure 4.3 shows the distribution of F_1-Normalized-Position training values for /i/ and /ɔ/. The values[2] for /i/ are generally lower than for /ɔ/ since /i/ generally has a lower F_1 than /ɔ/.

2. **F_1-Movement:** The movement of energy in the F_1 region is approximated by the slope of the best linear fit to the high/low-spectral-concentration over the middle 80% of the phone region:

$$\frac{N \sum_{t=t_{.1}}^{t_{.9}} t\, S_{F_1}(t) - \left(\sum_{t=t_{.1}}^{t_{.9}} t\right)\left(\sum_{t=t_{.1}}^{t_{.9}} S_{F_1}(t)\right)}{N \sum_{t=t_{.1}}^{t_{.9}} S_{F_1}(t)^2 - \left(\sum_{t=t_{.1}}^{t_{.9}} S_{F_1}(t)\right)^2}$$

where N is the number of samples from $t_{.1}$ to $t_{.9}$. This feature indicates how the energy below 1 kHz has shifted in frequency over the duration of a phone and is useful in discriminating between phones in which the average formant values are similar, but the amount of formant motion is different. Figure 4.4 shows the distribution of F_1-Movement training values for /aʸ/ and /ɔ/. The slope of /aʸ/ is more negative than the slope of /ɔ/ since F_1 falls in /aʸ/, but is approximately constant in /ɔ/.

[2] There are no units because spectral concentration parameters are normalized.

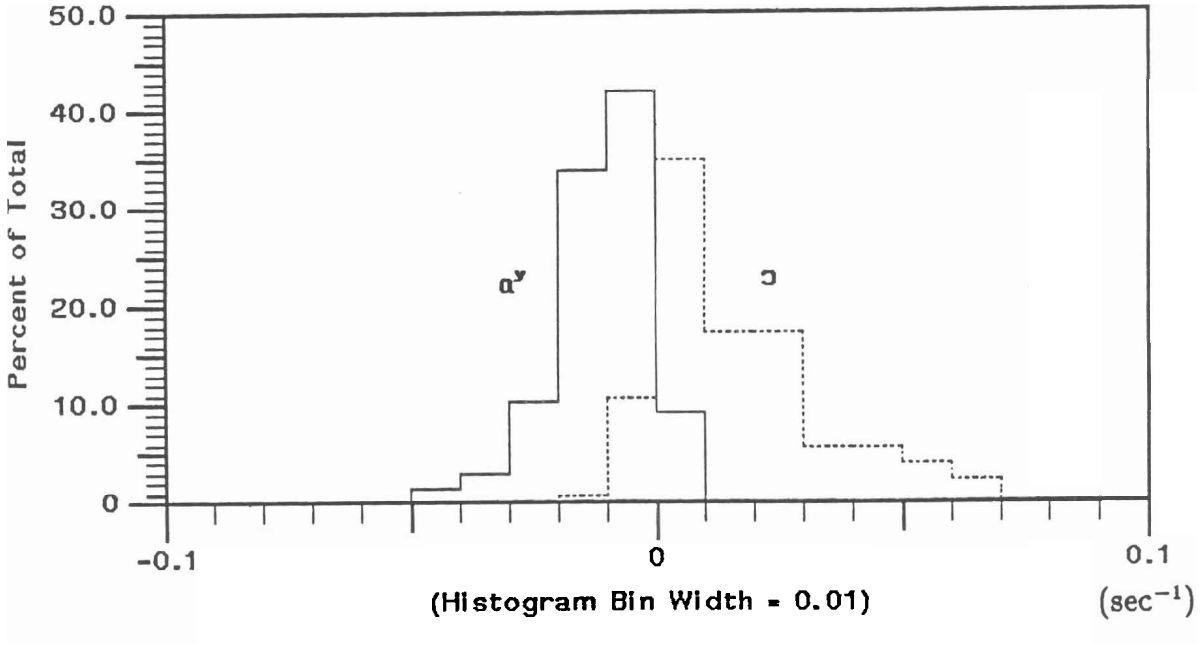


Figure 4.4: Distribution of F_1-Movement for /aʸ/ and /ɔ/

3. **F_2 Normalized Position:** This feature is the average value of the front/back-spectral-concentration parameter over the middle 50% of the phone region:

$$\frac{\sum_{t=t_{.25}}^{t_{.75}} S_{F_2}(t)}{t_{.75} - t_{.25}}$$

where $S_{F_2}(t)$ is the front/back-spectral-concentration at time t. This feature is most useful for identification of vocalic phones and indicates whether a phone is more like a front vowel or a back vowel. Figure 4.5 shows the distribution of F_2-Normalized-Position training values for /i/ and /ɔ/. The values for /i/ are generally higher than for /ɔ/ since /i/ has a higher F_2 than /ɔ/.

4. **F_2-Movement:** The movement of energy in the F_2 region is approximated by the slope of the best linear fit to the front/back-spectral-concentration over the middle 80% of the phone region:

$$\frac{N \sum_{t=t_{.1}}^{t_{.9}} t\, S_{F_2}(t) - \left(\sum_{t=t_{.1}}^{t_{.9}} t\right)\left(\sum_{t=t_{.1}}^{t_{.9}} S_{F_2}(t)\right)}{N \sum_{t=t_{.1}}^{t_{.9}} S_{F_2}(t)^2 - \left(\sum_{t=t_{.1}}^{t_{.9}} S_{F_2}(t)\right)^2}$$

This feature indicates how the energy in the range of F_2 has shifted in frequency over the duration of a phone. As with F_1-Movement, this feature is

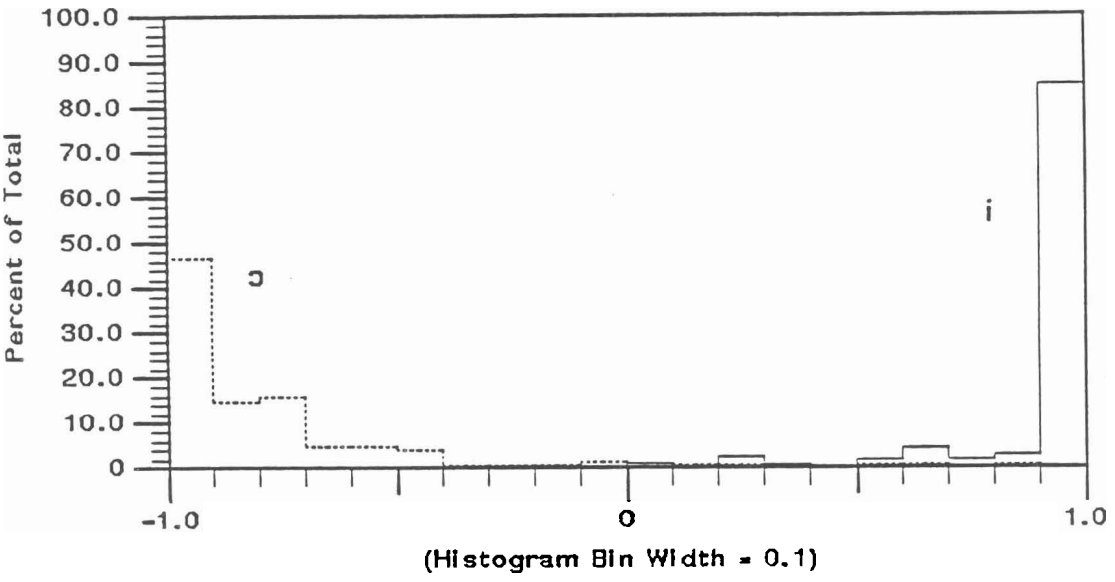


Figure 4.5: Distribution of F_2-Normalized-Position for /i/ and /ɔ/

useful in discriminating between phone pairs like /aʸ/ and /ɔ/. Figure 4.6 shows the distribution of F_2-Movement training values for /aʸ/ and /ɔ/. The slope of /aʸ/ is more positive than the slope of /ɔ/ since F_2 rises for /aʸ/ but is relatively constant for /ɔ/.

5. **/r/-Possibility:** This feature, which is useful in detecting /r/'s, is computed as the average value of the /r/-spectral-concentration over the middle 50% of the phone region:

$$\frac{\sum_{t=t_{.25}}^{t_{.75}} S_R(t)}{t_{.75} - t_{.25}}$$

where $S_R(t)$ is the /r/-spectral-concentration at time t. This feature indicates how much energy in the region of F_3 is below 2200 Hz, relative to energy above 2200 Hz. Figure 4.7 shows the distribution of /r/-Possibility training values for /r/ and /eʸ/. The values for /r/ are much lower than for /eʸ/ due to the lowering of F_3. The distribution includes pre-, post-, and intervocalic /r/'s. Better results should be obtained by computing the average value of /r/-spectral-concentration over different regions for the different allophones of /r/. That is, the feature for prevocalic /r/'s should be computed during the

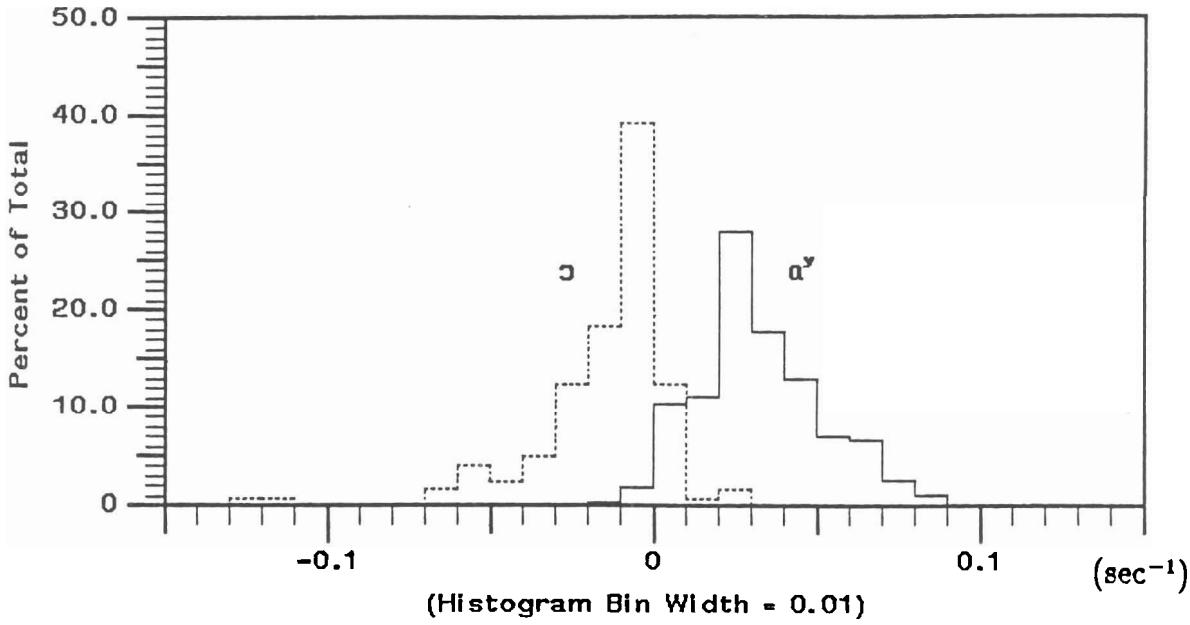


Figure 4.6: Distribution of F_2-Movement for /aʸ/ and /ɔ/

beginning portion of the phone, and the feature for postvocalic /r/'s should be computed during the final portion of the phone.

6. **Nasal-Possibility:** This feature is the average value of the difference in (log) energy computed with a passband of 100 to 350 Hz and (log) energy computed with a passband of 350 to 850 Hz:

$$\frac{\sum_{t=t_{.25}}^{t_{.75}} E_{100-350}(t) - E_{350-850}(t)}{t_{.75} - t_{.25}}$$

In each energy computation, 50 Hz tapers on the trapezoidal frequency window were used. This feature captures the presence of the low resonance around 300 Hz which is characteristic of nasal murmurs (Fujimura, 1962). Figure 4.8 shows the distribution of Nasal-Possibility training values for /n/ and /ɔ/. The values for nasal consonants are generally larger than for non-nasals due to the presence of energy from the low resonance in the lower band.

7. **Onset-Rate:** This feature measures the maximum change in energy from 1000 Hz to 7000 Hz within 20 msec of the beginning of a phone:

$$\max_{i=t_b-d}^{t_b+d} [E_{1000-7000}(i) - E_{1000-7000}(i-2)]$$

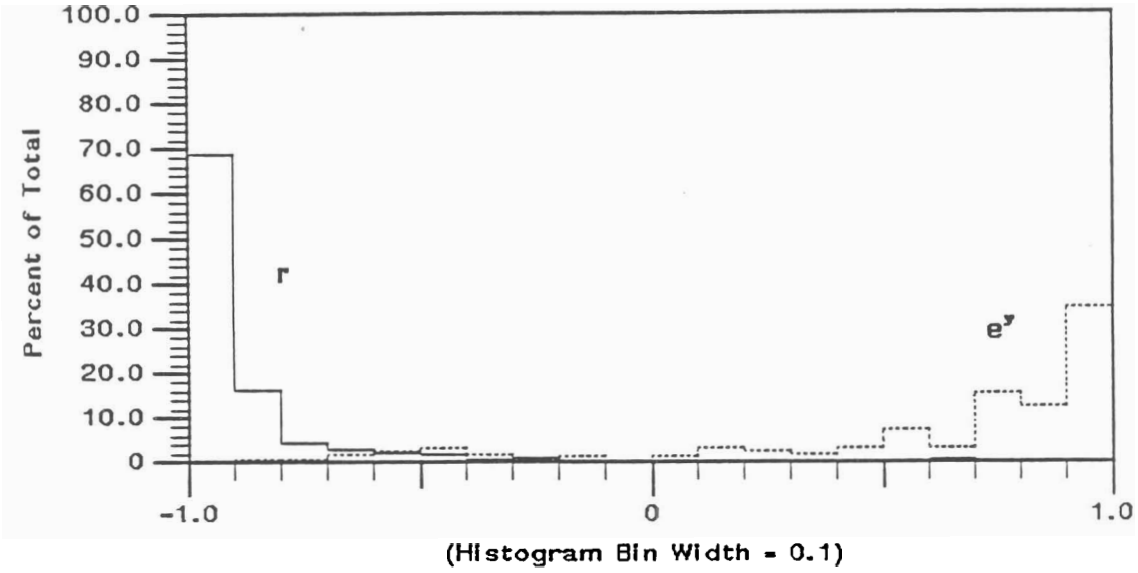


Figure 4.7: Distribution of /r/-Possibility for /r/ and /e/

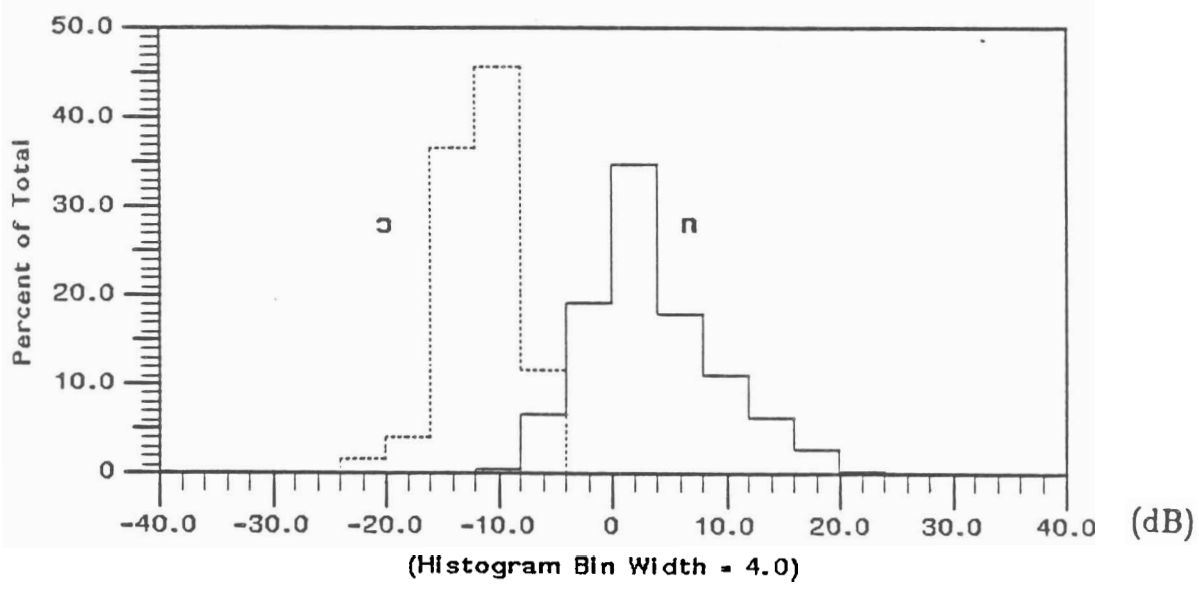


Figure 4.8: Distribution of Nasal-Possibility for /n/ and /ɔ/

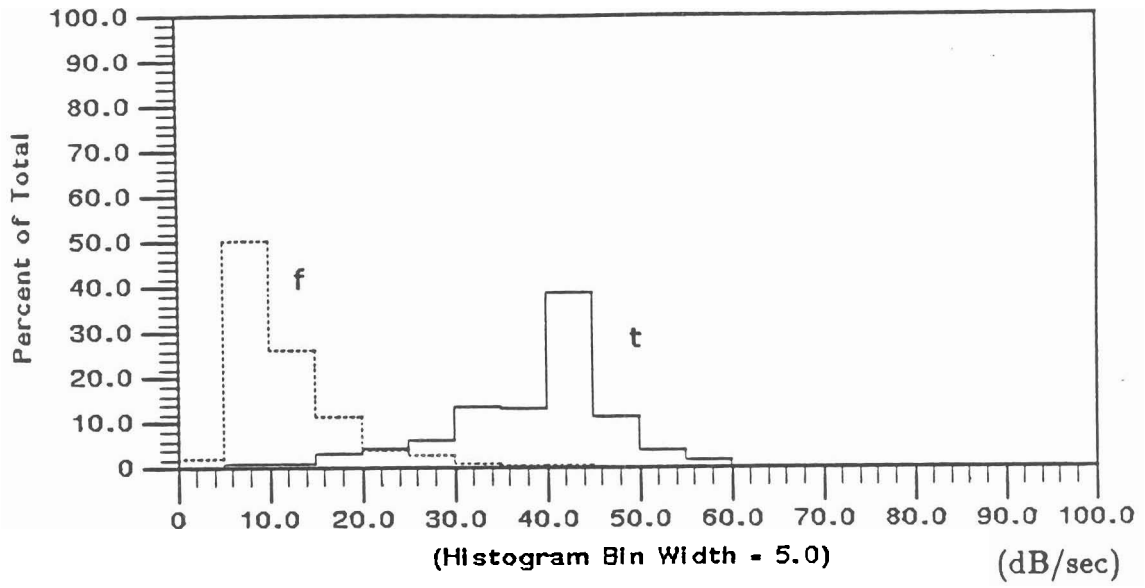


Figure 4.9: Distribution of Onset-Rate for [t] and [f]

where i is the sample number, t_b is the sample at the time the phone begins, and d is the number of samples in 20 msec. In order to capture rapid transitions, $E_{1000-7000}$ was computed every msec from the short time Fourier transform using a Hamming window width of 2 msec. This feature is particularly useful for discriminating stops from fricatives because stop onsets are generally much more rapid than fricative onsets. Figure 4.9 shows the distribution of Onset-Rate training values for [t] and [f]. The Onset-Rate is larger for the release of [t] than for [f] as expected.

8. **Spectral-Offset-Location:** This feature indicates the location of the first major spectral dip in the cepstrally smoothed spectrum 30% through the duration of the phone. The time at which this feature is computed was chosen empirically and was motivated by the task for which the feature was designed. This feature was initially designed to discriminate between the /aʸ/ in "five" and /ɔ/ in "four." The spectrogram in Figure 4.10 contains the word sequence "five four," and the location of the /aʸ/ and /ɔ/ are indicated below. Comparing the /ɔ/ in "four" with the /aʸ/ in "five," we note that the strik-

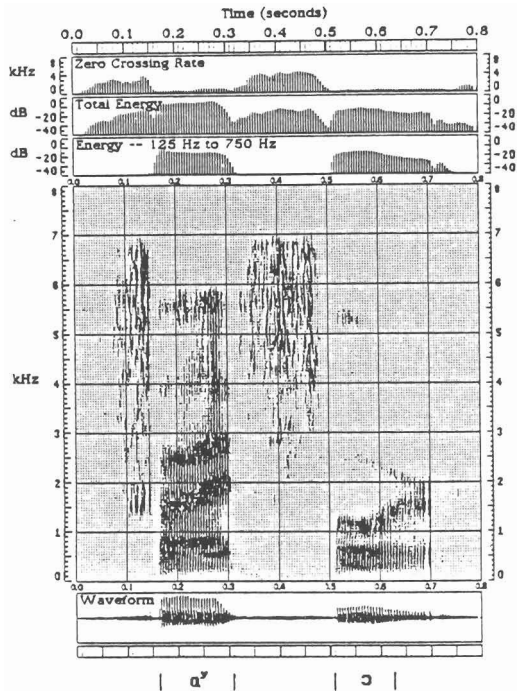


Figure 4.10: Spectrogram of "five four"

ing differences are in the location of F_2 and lack of energy between F_2 and F_3. This feature tries to capture the location of the upper edge of F_2 to differentiate between the two sounds. Figure 4.11 shows the distribution of Spectral-Offset-Location training values for /aʸ/ and /ɔ/.

9. **High-Frequency-Energy-Change:** The change in high frequency energy is computed as the slope of the best linear fit to the energy in the 4500-7800 Hz band over the middle 80% of the phone region:

$$\frac{N \sum_{t=t_{.1}}^{t_{.9}} t\, H(t) - \left(\sum_{t=t_{.1}}^{t_{.9}} t\right)\left(\sum_{t=t_{.1}}^{t_{.9}} H(t)\right)}{N \sum_{t=t_{.1}}^{t_{.9}} H(t)^2 - \left(\sum_{t=t_{.1}}^{t_{.9}} H(t)\right)^2}$$

where H(t) is the value of high frequency energy at time t. This parameter is intended to help differentiate between fricatives and stop releases. Fricatives are relatively stable over their duration; in contrast, unvoiced stop releases generally have a strong onset followed by aspiration which weakens over the duration of the phone. Thus the slope is expected to be more negative for a stop release, such as [t], than for a fricative such as [s]. Figure 4.12 shows

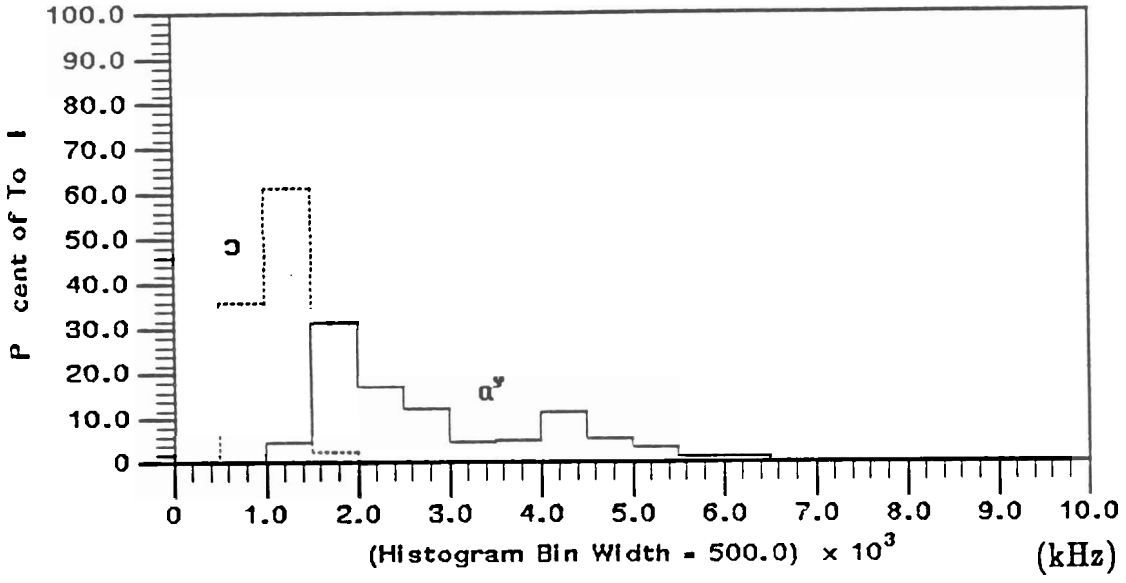


Figure 4.11: Distribution of Spectral-Offset-Location for /aʸ/ and /ɔ/

the distribution of High-Frequency-Energy-Change training values for [s] and the release of [t]. The values for [s] center around 0, since fricatives generally do not weaken; the values for [t] are generally negative since high frequency energy usually decreases after the release in a stop.

4.2 Scoring Word Hypotheses

The task of assigning a score to each phone in a word lattice may be approached as a discrimination and/or identification problem. When viewed as a discrimination task, a binary discrimination is performed between each pair of competitor phones. The results are then used to assign a score to each phone indicating how good the hypothesized phone is relative to its competitors. When viewed as an identification task, each phone is assigned a score of how good it is, independent of the values of the other phones.

Since the lexical component has already reduced the competitors to a small subset of words, discrimination between the remaining competitors should give better performance than trying to identify a phone from all possible phones. In general,

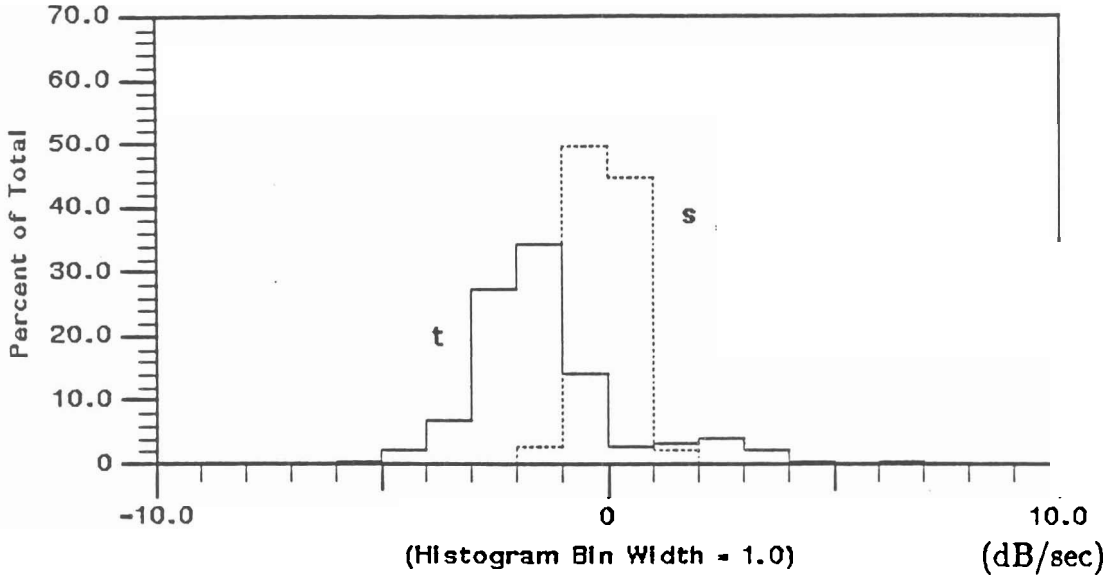


Figure 4.12: Distribution of High-Frequency-Energy-Change for [t] and [s]

it is easier to identify an object from a small group of objects by comparing it to each of the possible alternatives (discrimination) using knowledge about how they previously compared than it is to identify an object based only on knowledge of the characteristics of each object (identification). However, defining competitor phones in a meaningful way is difficult because phone boundaries are not always aligned. In contrast, identification does not require competitors to be defined. So although we expected discrimination to perform better, we explored both methods for verification. The same algorithm was used to compute word scores from the identification or discrimination scores, and the results from the two methods were compared. The identification and discrimination scores represent the score of a phone based upon information from particular detailed feature. These conditional phone scores were combined in this algorithm to produce phone scores, and the phone scores were combined to produce word scores.

Each approach required a priori knowledge of the distribution of the features used. This knowledge was provided from values computed on a set of training utterances described in Appendix A.

4.2.1 Phone Scores Based on Identification

In the identification task, observed values of the feature vector for a hypothesized phone and training values for the features are used to compute phone scores. The scores reflect the probability of a phone occurring given the observed feature values. In particular, the unnormalized score of phone i based on information from feature k, $S_u(p_i|f_k)$, was computed as:

$$S_u(p_i|f_k) = \Pr(p_i|f_k) = \frac{\Pr(f_k|p_i)\Pr(p_i)}{\Pr(f_k)}$$

where phone i is represented as p_i and feature k is represented as f_k. Assuming that each phone is equally likely, $\Pr(p_i)$ is a constant and can be ignored. Similarly, $\Pr(f_k)$ is the same for all phones being evaluated in the region and can also be ignored. Therefore, the score of phone i based upon information from feature k is proportional to $\Pr(f_k|p_i)$.

A k-nearest-neighbor estimate was used to compute $\Pr(f_k|p_i)$. Since the k-nearest-neighbor estimate is a function of the window width required to capture k samples surrounding a point, the estimate is a function of the range of each of the features. To permit information from different features to be combined, each estimate was normalized by the range, R, of the observed values for the feature over all phones. The normalized score of phone i based on information from feature k, $S(p_i|f_k)$, was thus computed as:

$$S(p_i|f_k) = \frac{\Pr(f_k|p_i)}{R}$$

4.2.2 Phone Scores Based on Discrimination

In the discrimination task, the score for each phone based upon a particular feature is a function of how well the phone compares to each of the competitor phones according to that feature. Discrimination of the phones in the word lattice produced by the lexical access component was expected to give better results than

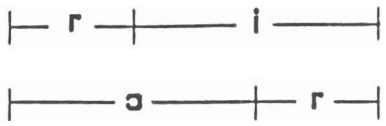


Figure 4.13: Alignment of /ri/ compared to /ɔr/

identification. This is because the recognition model used sequential constraints to remove unlikely word candidates from further consideration. The remaining word candidates are a small subset of all possible word candidates which can be discriminated using explicit knowledge of the limited number of competitors. Since each remaining phone candidate within a broad class segment matches the segment well and the phone candidates also match the initial features characterizing the broad class segment well, the remaining phones are similar to each other. Thus fine discriminations must be made in order to accurately score how well each phone is realized relative to the other phone candidates.

To compute a discrimination score, the phone being scored was first compared to each of its competitors. However, competitors can be defined in many ways. Ideally, competitors should cover the same region of an utterance. Since lexical access is performed at a broad phonetic level, one or more phones may map into one broad class, depending upon the hypothesized word. For example, /ɔr/ and /ri/ may both map to the broad class "vowel." Since /r/ in intrinsically shorter than the adjacent vowel, the boundary between /ɔ/ and /r/ will be after the boundary between /r/ and /i/, as illustrated in Figure 4.13

As seen in the previous example, the phones will not always line up, and a choice must be made as to whether or not all competitor phones should be forced to have the same endpoints. By requiring all competitors to have the same endpoints, either the boundaries of a phone will not be accurate, or the recognition unit will be composed of a variable number of phones. Consequently, separate acoustic features have to be developed which look for characteristics of a phone within a region. For

example, the vowel portion of the broad phonetic sequence "weak-fricative vowel silence" may map to /ri/, /rieᵞ/, or /ɔr/. A separate feature for /r/ would then be required for each sequence. A problem that results from this approach is that as the vocabulary grows, the number of possible sequences grows. That is, a new word-initial(final) phone sequence may form new sequences with word-final(initial) phone sequences in the vocabulary when adjacent phones can geminate. In contrast, choosing the phone as the unit to be scored eliminates the need for separate feature detectors; only detectors for the phones in the vocabulary are needed.

Competitors were defined as any phone which overlaps in time with the phone being scored. The probability of phone i relative to each competitor phone j was computed using Bayes' Rule under the assumption that each phone is equally likely:

$$Pr_j(p_i|f_k) = \frac{\Pr(f_k|p_i)}{\Pr(f_k|p_i) + \Pr(f_k|p_j)}$$

As in identification, $\Pr(f_k|p_j)$ was computed using a k-nearest-neighbor estimate. Normalization by the feature value range was not performed since the score is computed as a ratio.

The score for phone i based on information from feature k, $S(p_i|f_k)$, was then computed as the average of how likely it is that phone i is the underlying phone relative to each competitor phone:

$$S(p_i|f_k) = \frac{\sum_{j=1}^{J} \Pr_j(p_i|f_k)}{J}$$

where J is the number of competitor phones. Thus the discrimination score is based on how phone i compares only to the competitor phones. This can be contrasted to the identification task where phone i is scored based upon how well it matches previous observations of the phone.

4.2.3 Computation of Phone and Word Scores

The score of a phone is a function of the set of scores for that phone conditionalized on different features. We can think of these scores as reflecting how strongly

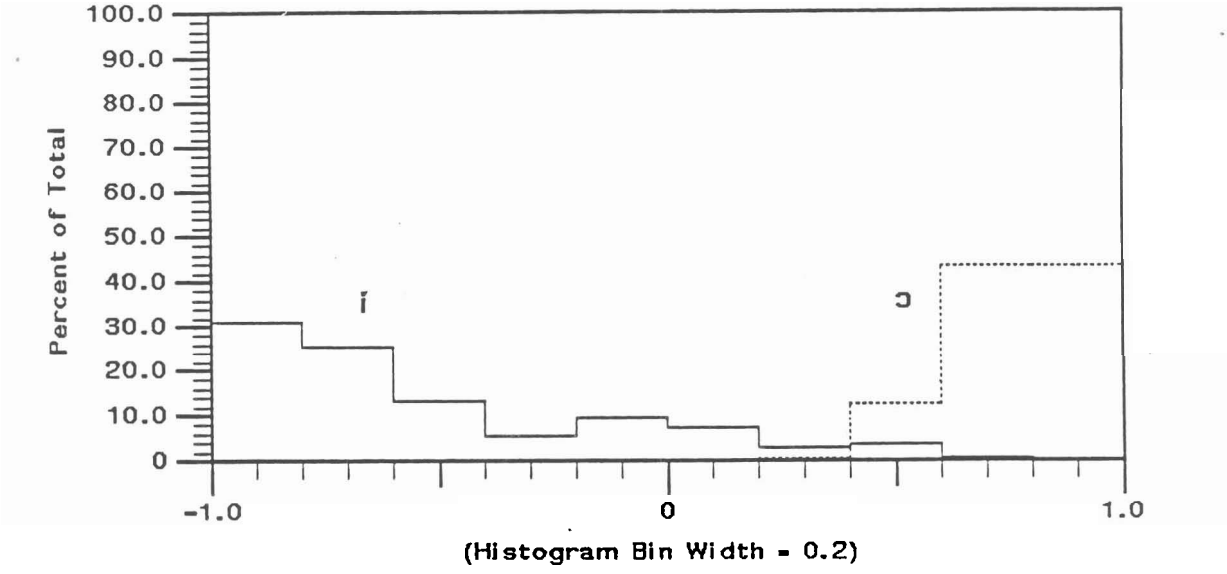


Figure 4.14: Percentage Histogram of /i/ and /ɔ/ Divided into 10 Bins

each feature indicates that the underlying phone is the hypothesized phone. Each feature provides input to the decision process, weighted by the quality of that feature in identifying the phone.

The weights for each phone were computed by measuring how well each feature identifies the phone from possible competitor phones. The next few paragraphs describe how the quality of a feature for identifying a phone from its competitors is computed.

The range of a feature, as observed over all phones was first divided into N bins. In this study, N was empirically chosen to be 10 after examining the number of training tokens per phone. With too few bins, the quantized counts would show little variation. With too many bins, most of the bins would contain either 0 or 1 samples, and the counts would again be meaningless. The percentage of samples which fall into each bin was then computed for each phone. Figure 4.14 illustrates the binning of the F_1-Normalized-Position feature for /i/ and /ɔ/.

For phone pairs with very dissimilar distributions, as in Figure 4.14, many of the bins contain samples primarily from one type of phone. To determine the

dissimilarity of a pair of phones, the percentage of samples in each bin was compared. The dissimilarity, $d_k^{ij}(b)$, was measured as:

$$d_k^{ij}(b) = \frac{|s_{bi} - s_{bj}|}{s_{bi} + s_{bj}}$$

where s_{bi} is the percentage of phone i samples in bin b and s_{bj} is the percentage of phone j samples in bin b. From this equation, we note that when only one type of phone is present in a bin, the maximum bin dissimilarity score of 1 was assigned. If the values in a bin for the two phones being compared are equal, then the minimum bin dissimilarity score of 0 was assigned.

The quality, q_k^{ij}, of feature k for discriminating between a pair of phones, phone i and phone j, is the average of the bin dissimilarities:

$$q_k^{ij} = \sum_{b=1}^{B} \frac{d_k^{ij}(b)}{B}$$

Bins in which no phones are present were ignored. B thus is equal to the number of bins with at least one sample. The quality of a feature for identifying a phone, q_k^i, is the average quality of the feature for discriminating between each possible phone pair:

$$q_k^i = \frac{\sum_{j=1}^{J} q_k^{ij}}{J}$$

where J is the number of phones against which phone i may be compared.

The quality of a feature is used to weight the input of each feature geometrically:

$$Score(p_i) = \frac{\sum_k q_k^i \log S(p_i|f_k)}{\sum_k q_k^i}$$

This weighting was used because we wanted to emphasize the features which are good in identifying a phone and minimize the effect of measurement noise that nonrobust features add. That is, if only a few features are useful in the identification of a phone, input from the features which are not meaningful should be minimized.

The score for each word was computed as a function of the score of each of the component phones. Each phone was weighted by its duration in order to normalize

for a varying number of phones in each path in the word lattice. For example, /θəri/ and /faʸ/ may both be word candidates over the same region of speech. Since the number of phones in the two words is different, the score for each phone cannot simply be added; the path scores must be normalized to remove the effect that one path contains twice as many phones.

There are several alternatives for computing normalized word or path scores. The phone scores could be summed and then normalized by the number of phones. This method has the advantage that long phones are not given undue weight relative to short phones. It also has a disadvantage which is illustrated by the /θəri/-/faʸ/ example. If /θəri/ is normalized by the four phones composing it, then the weak fricative /θ/ receives a weight of .25. However, in /faʸ/, the weak fricative /f/ receives a weight of .5. Thus the weight given a phone is dependent on the number of other phones in a word or sentence. Furthermore, when alternate pronunciations of the same word are compared, for example, /θəri/ and /θri/, then the /θ/ in both pronunciations should be given equal weight. By normalizing by the number of phones, the /θ/ in /θri/ receives more weight. Therefore, normalization by the number of phones has the undesirable effect that depending upon the word being verified, a phone may have a variable amount of input into the decision process. The chosen method of normalization, weighting by duration, avoids this problem, although it does have the disadvantage that short segments are given less weight. Normalizing the segments by duration, the word scores were computed as:

$$\text{Score}_{word} = \sum_{i} \text{Score}(p_i)\text{D}(p_i)$$

where $\text{D}(p_i)$ is the duration of phone i.

The simplifying assumption that each feature is independent was used in the scoring algorithm. This assumption was made to help insure that enough training samples were available to get good estimates since the number of samples required increases exponentially with the number of features (Duda and Hart, 1973). This technique ignores potential multivariate information, and correlations between de-

pendent features cannot be used. For example, two sets of data may be well separated in a two-dimensional feature space, but when the data is collapsed to a one dimensional feature space, the two sets of data may overlap and cannot be separated as well. Thus results obtained using this simple technique provide a bottom line on results which can be expected if a more sophisticated verification algorithm is used.

4.3 Computation of the Ideal Word Lattice and Phone Boundaries

The task of evaluating the feature set and scoring algorithm was structured such that the performance of the verifier could be evaluated independently from the performance of the earlier components. An "ideal" word lattice, free from segmentation errors, served as input to the verifier. By using error-free input, methods for handling earlier segmentation errors, such as verification or definition of each boundary were unnecessary. Instead, the study focused on the selection of features for discrimination between similar phones and on the utility of a phone representation for evaluating word hypotheses. The results of this study serve to indicate the viability of using acoustic-phonetic features for verification.

To compute phone scores, as outlined in the previous section, the phonetic transcription of each word hypothesis and the location of phone boundaries must be known. A word lattice contains information about the word endpoints and segment boundaries. A simple procedure was used to locate the phone boundaries from the information in the word lattice. In this section, the procedures for computing the ideal word lattice and locating phone boundaries are outlined.

4.3.1 Computation of the Ideal Word Lattice

The ideal word lattice was derived by mapping the hand labeled phonetic tran-

scription into a broad phonetic segmentation and then hypothesizing words by matching the words in the lexicon to sections of the broad segmentation. Rules were used in the mapping procedure to adjust boundaries and account for transcription labels which did not map directly into a broad phonetic class. To simulate the segmentation which would be produced by the broad phonetic classifier within a word, adjacent phones belonging to the same broad class were represented by a single broad phonetic label. For example, the /z/ in "zero" maps to "strong-fricative" and the /ɪ/, /r/, and /oʷ/ all map to "vowel." Thus in the broad phonetic representation of "zero," /ɪ/, /r/, and /oʷ/ are represented by one vowel segment and "zero" is represented as "strong-fricative vowel." Acoustic gemination was not accounted for in this mapping since adjacent phones which belong to the same broad class but occur in successive words are represented by two separate segments.

The mapping procedure modified the endpoints of the derived broad phonetic transcription to be different than the endpoints in the phonetic segmentation when sounds did not directly map into a broad phonetic class. This occurred when noise, glottalization, voicebar, aspiration, or epenthetic silence were encountered. In these cases, the segment was arbitrarily divided evenly between the adjacent labels. This procedure produced an idealized segmentation for matching with the words in the lexicon.

In continuous speech, the location of word endpoints are unknown a priori. Words and their corresponding endpoints were hypothesized by matching each word in the lexicon against sections of the broad phonetic segmentation. All words matching a section of the broad phonetic transcription were collected to produce an "ideal" word lattice. Each of the words in the lattice contained the phonetic transcription of the word and the sequence of broad phonetic labels with associated endpoints.

The depth of the ideal word lattice, that is, the number of words in the lattice divided by the number of digits in the digit string, is statistically characterized in Figure 4.15. The average depth, which was computed before path and allophonic

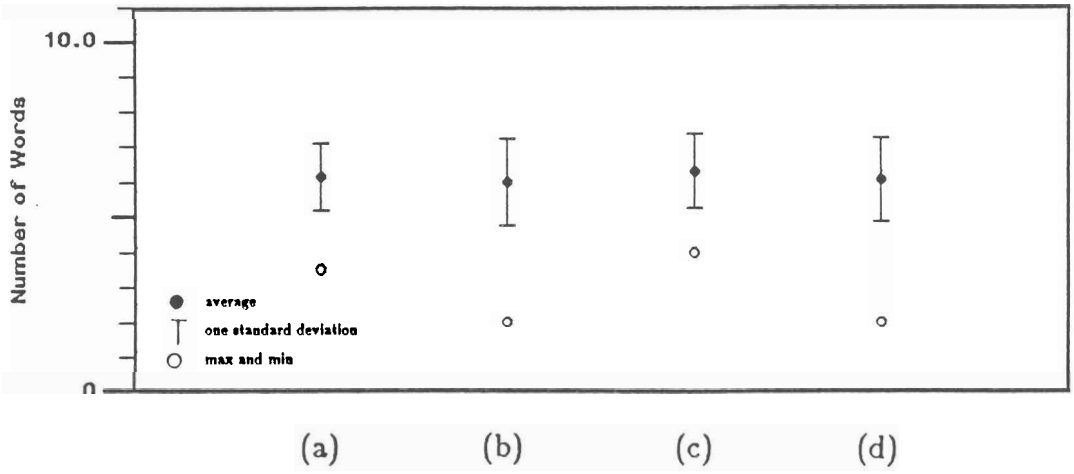


Figure 4.15: Number of Word Candidates in the Ideal Word Lattice per Word in Digit String (a) training utterances by training speakers (b) new utterances by training speakers (c) training utterances by new speakers (d) new utterances by new speakers

constraints were applied, is about 6 words/digit. This depth corresponds to a sequential constraint pruning threshold of approximately -0.5. Thus, there is much room for improvement in the performance of the broad phonetic classifier and lexical component before it approximates the ideal case.

4.3.2 Computation of Phone Boundaries

From the information contained in each hypothesized word, the verification component determined any phone boundaries not specified by the broad phonetic segmentation. It is necessary to find boundaries within a phonetic segment when multiple phones have been mapped to one segment. When the phonetic transcription of a hypothesized word contained an intervocalic /r/, an r-detector was used to locate the boundaries of /r/. The intervocalic /r/ detector first located the time, t_{min}, at which the median smoothed /r/-spectral-concentration parameter is

a minimum. Anchoring from this point, the /r/-spectral-concentration parameter was searched outward in both directions, as in broad phonetic classification, until the parameter rose to 30% of the difference between the value at t_{min} and the minimum of the local maxima on each side. The front/back-spectral-concentration parameter was also used to locate the /r/ boundary when in a front vowel context. This is because the /r/-spectral-concentration parameter tends to remain low for a longer time due to the higher frequency location of F_2. The front/back-spectral-concentration parameter was searched from t_{min} until it rose to 30% of the difference between its value at t_{min} and the minimum of the local maxima on each side. The minimum distance from t_{min} to the computed edge for the /r/- and front/back-spectral-concentration parameters were defined as the /r/ boundaries. When more than one phone mapped to a broad phonetic segment and an intervocalic /r/ was not present, a default weighting, which assigns /w/ and /r/ half the duration of the other phones, was used to divide the time among the phones mapping to the segment.

4.4 Evaluation

The performance of the verifier was evaluated in terms of the word error rate and the rank of each phone in the correct path. Because the task is continuous speech, the words were evaluated subject to path constraints. That is, the "best" words were the string of words which formed the best scoring path through the word lattice. Separate evaluation of each word relative to a hand labeled orthographic transcription was not used because the meaning of comparing competitor words which have different endpoints is unclear.

The word error rate was computed by observing how often the words comprising the best word string did not match the words comprising the correct word string. With this method, a word could be the best scoring word over a region but may

not be in the best path. Insertions, deletions, substitutions and matches were computed using a 50% overlap criteria similar to that used to evaluate the broad phonetic classifier. Thus, if at least 50% of each of two words in the best word string is covered by one correct word, then an insertion is said to have occurred. Pauses between words were not included in the computation of word errors.

The best-scoring path through the word network was found using a depth-first search (Winston, 1984) without any constraints on the number of words in the digit string. The use of an exhaustive search algorithm insures that the system finds the best answer from the information it is given. Thus the effectiveness of the set of detailed features combined with the scoring algorithm was evaluated independently of any heuristics which could be used to reduce the search time.

4.4.1 Discrimination vs Identification

The discrimination and identification scoring methods were compared on a subset of the training data. The word error rate using the identification method was 3.6% and the word error rate using the discrimination method was 2.0%. As discussed earlier, the discrimination scoring method was expected to give better results because the word candidates had been reduced to a set where fine differences between competitor phones existed; thus a discrimination paradigm was more appropriate. Since the results bear out this expectation, the rest of the evaluations were performed using the discrimination scoring method.

4.4.2 Word Errors

Table 4.1 shows the word error rates for different testing conditions. Each insertion, deletion, or substitution was counted as an error. The error rate of 1.5% on the training utterances by training speakers illustrates the power of using a few carefully selected acoustic features combined with statistical measures to estimate the goodness of a phone. The error rate for training speakers on new utterances is

Table 4.1: Word Error Rates

Utterances	Speakers	# of Speakers	# of Digits	Word Error Rate
training	training	6	1365	1.5%
new	training	4	1126	5.0%
training	new	3	364	4.9%
new	new	4	893	5.3%

approximately the same as the error rate for new speakers on new utterances; this indicates that an acoustic-phonetic approach is potentially speaker-independent. The error rate for training utterances and new utterances spoken by new speakers is approximately 5%. This shows that for new speakers, the system can handle new utterances about as well as the utterances it was trained on.

A more detailed analysis of the errors in all corpora reveals that many of the errors were due to male/female differences. Some of these errors are listed in Table 4.2. The most striking and consistent error is the confusions of "four" and "five". All 16 cases in which "five" was mistakenly labeled as "four" occurred in speech by males. Eighteen of the 19 cases in which "four" was confused as "five" occurred in speech spoken by females.

Considering the acoustic differences between "four" and "five," and the differences between male and female speech, these errors can be attributed to selecting features which are not independent of male/female differences. The first phone in both "four" and "five" is /f/. Since both /ɔ/ and the initial portion of /aʸ/ are low back vowels, the coarticulation effects on /f/ due to the following vowel are approximately the same. Hence, the /f/ is similar in both words, and the main difference between the two words lies in the vocalic portion. In the vocalic portion, F_2 rises in both "four" and "five": in "four" it rises for the production of /r/, and in "five"

Table 4.2: Sample Male and Female Word Errors

Digit	Recognized as	# Males	# Females
four	five	1	18
five	four	16	0
three	four	0	7
two	zero	0	9
zero	seven	3	13

it rises during the latter portion of the /aʸ/. One of the primary differences is the higher initial location of F_2 in /aʸ/ than in /ɔ/. However, for the same vowel, the location of F_2 varies among speakers. It is generally higher in frequency for female speech than male speech (Peterson and Barney, 1952), since females have a shorter vocal tract length. The Spectral-Offset-Location was designed to be sensitive to differences in the initial location of F_2 in /aʸ/ and /ɔ/. Considering this sensitivity, the errors of labeling the /ɔ/ in "four" as an /aʸ/ in female speech, and the /aʸ/ in "five" as an /ɔ/ in male speech, are reasonable. A better, speaker-independent feature may be to compare the Spectral-Offset-Location relative to the location of F_3, since a larger dip in the spectrum is observed in /ɔ/ than in /aʸ/.

To obtain an idea of the robustness of the verification scores, the score of the correct word relative to the score of the other word candidates was examined. In particular, the score of the top candidate was compared to the score of the second best candidate when the top candidate was correct. When the top candidate was incorrect, its score was compared to the score of the correct word. Figure 4.16 illustrates this for each of the test sets. Note that the difference in word scores is generally small when an incorrect word is the best scoring word, and that the difference has a large range when the correct word is the best scoring word. In a recognition system, this information could be used to reject the utterance when

Table 4.3: Phone Rank in Correct Words

Speakers	Utterances	Position			
		0	1	2	3
Training	Training	.93	.99	1.00	1.00
Training	New	.86	.98	1.00	1.00
New	Training	.90	.99	.99	1.00
New	New	.86	.98	.99	1.00

the difference in word scores is small, and the speaker could be asked to repeat the utterance. Alternatively, this information could be used to identify words which do not score much better than their competitors, and finer discriminations could be performed on these words.

4.4.3 Phone Errors

The rank of each phone in the correct word was also used to evaluate the verifier. These results are shown in Table 4.3 for the test sets. Note that for new sentences by both the training speakers and the new speakers, the correct phone is in the top position at least 86% of the time and within the top two candidates at least 98% of the time. This similarity in rank indicates the potential speaker-independence of using acoustic features for verification. As expected, the largest percentage of phones were in the top position when the system was tested on the test set composed of training utterances by training speakers. However, the top two ranking candidates include the correct phone at least 98% of the time on all corpora. These results indicate the viability of performing verification at the phone level using acoustic-phonetic features.

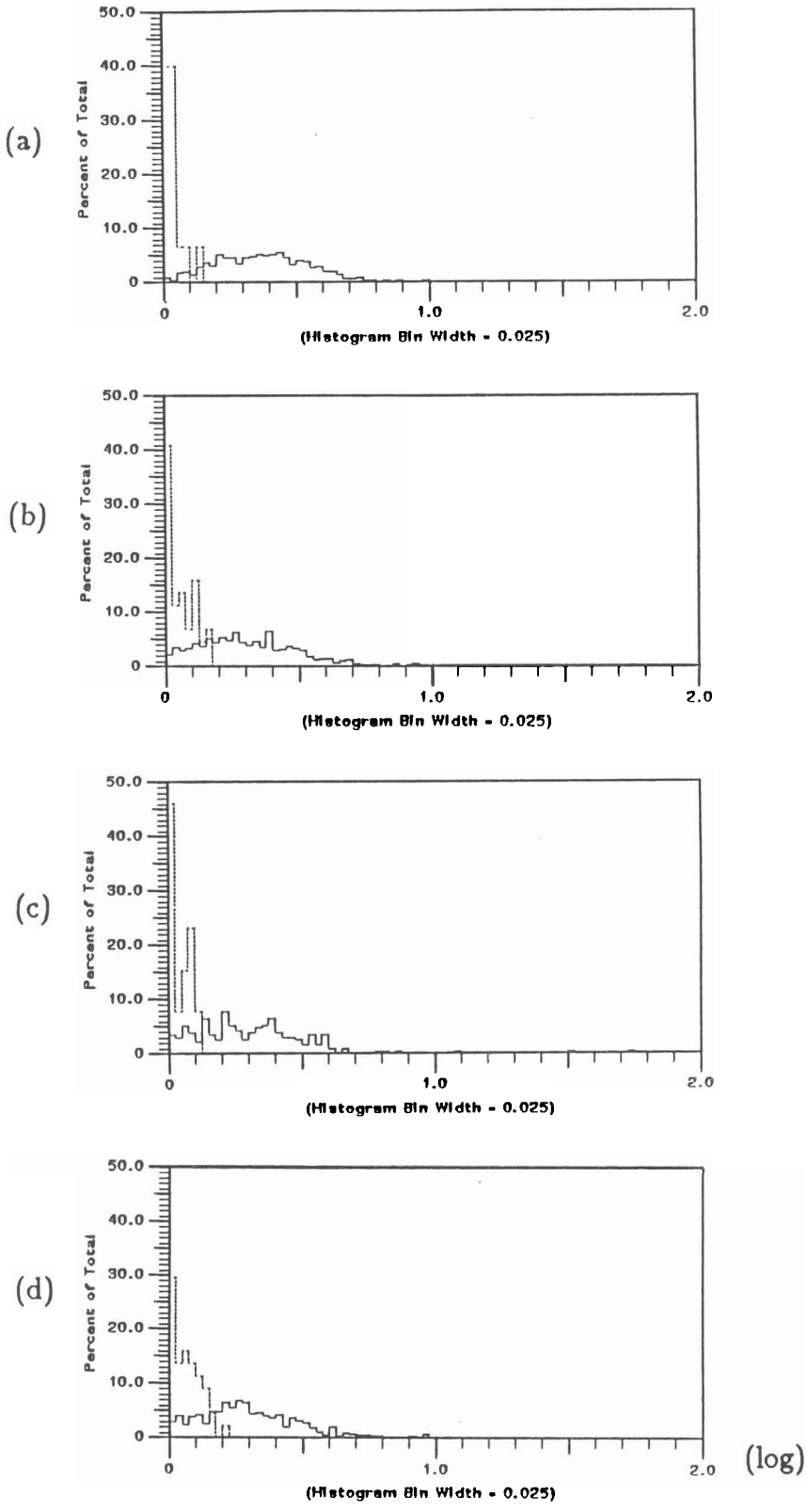


Figure 4.16: Difference in Word Scores between Correct Word and Next Best Word (solid line) and Best Word and Correct Word (dashed line) for: (a) training utterances by training speakers (b) new utterances by training speakers (c) training utterances by new speakers and (d) new utterances by new speakers

4.5 Chapter Summary

The main points of this chapter are:

- Words were scored based upon the value of phone features, demonstrating a potentially extendable scoring method.
- Use of a phone representation allows features to be defined over regions which are minimally affected by coarticulation and also allows a wider variety of characterizations of speech to be exploited.
- Better performance was achieved using discrimination between phone competitors than identification based purely on feature values.
- Verification using acoustic features is potentially speaker-independent.
- A small set of well chosen acoustic features is adequate for verification of phones in the digit vocabulary.

Chapter 5

Discussion

This chapter discusses the underlying assumptions and contributions of this thesis research, which was focused on examining how low-level acoustic-phonetic speech knowledge can be used in continuous speech recognition. Several underlying assumptions were made in this research. One of the basic assumptions was that an acoustic-phonetic approach is worth exploring. It was also assumed that speech can be represented as a sequence of sounds associated with regions of the speech signal, and that a phone is a good unit for representing speech. The following sections attempt to clarify the reasoning behind these assumptions and the choice of the digit vocabulary as a case study. Additionally, the merit of each component in the acoustic-phonetic recognition model is considered. The use of a preprocessor, in particular, a preprocessor based on acoustic-phonetics, in a recognition system is discussed. The reasons behind the choice of a simple control strategy for studying how detailed acoustic features can be used in verification is addressed. Computational requirements of a segment-based approach are also considered. Finally, the contributions of this thesis towards a better understanding of the use of acoustic-phonetic knowledge in speech recognition is outlined, and suggestions for future work are made.

5.1 Why an Acoustic-Phonetic Approach?

An acoustic-phonetic approach is appealing both intuitively and in its potential extendability to less restrictive recognition tasks. It is intuitively appealing because it provides a framework for describing speech sounds and coarticulation and also for applying such knowledge in a recognition system. Since the knowledge used by the system is specified using human knowledge, its application is often under explicit human control.

By incorporating speech knowledge–knowledge about the problem domain–the errors which the system produces are usually reasonable. That is, one can usually understand why the errors occurred. Therefore, the errors may be corrected by extending the knowledge. For example, when computing phone scores based on how likely the feature values indicate that a phone is the hypothesized phone, many of the errors appeared to result from differences between male and female speech. By choosing features which are less dependent on these differences, or by normalizing the values based upon speaker characteristics, these differences may be more readily handled.

In contrast, other approaches do not provide a way to explicitly handle such errors. These approaches incorporate speech knowledge only implicitly in the training data or by folding the information into the recognition algorithm. Thus the system must either be retrained or the recognition algorithm must be modified to "correct" errors which occur only in a subset of phones. In template matching, researchers have the option to increase the amount of training data and the number of reference templates used. However, this has the undesirable consequence that attempts to correct errors which occur with only one type of speech sound increase the number of templates for all recognition units. In addition, researchers must explicitly collect data containing the variation or else the new templates may be based upon only a few outliers, resulting in non-robust templates.

An acoustic-phonetic approach is also appealing in its potential extendability to

larger vocabularies and more complex task domains. The extendability results from the use of a phone based representation. The number of phones in any language is limited. English uses about only 40 phones; hence, the maximum possible number of diphones, or phone pairs, is limited to about 40^2. Furthermore, the number of consonant phone sequences within a syllable is finite. Studies have been conducted on the number of unique consonant phone sequences in a subset of commonly occurring English words. These studies show that as new words are added, new phone sequences occur, as expected. More importantly, these studies show that the possible number of allowable sequences within a syllable is limited. In particular, only about 70 syllable-initial and 130 syllable-final consonant sequences exist in English (e.g., Shipman and Zue, 1982). Consequently, the maximum number of contexts in which a phone can occur is much less than the number of words in English. Since each word can be represented as a phone sequence in an acoustic-phonetic approach, and since the number of recognition units is independent of the number of words in such a representation, an acoustic-phonetic approach is potentially extendable.

A considerable investment is required to develop the knowledge needed to characterize phones in different environments for a phonetically based system. However, once this investment has been made, the vocabulary should be extendable simply by adding one or more phonetic representations for each new word. In contrast, a word-based system must train each new word, possibly in each environment in which it could occur. Thus, we believe that the investment needed to gather knowledge for the development of acoustic-phonetically based systems has large potential payoffs by providing a framework for exploiting speech knowledge and for developing less restrictive speech recognition systems.

5.2 Why the Use of Digits as a Case Study?

Selection of the digits as a case study for exploring constraints in continuous

speech recognition is the result of our attempt to demonstrate the model, and at the same time to keep the problem manageable. Admittedly, the digit vocabulary is limited in many ways. For example, syntax and semantics are not applicable to random-length digit strings. In addition, stress is not of primary importance because the amount of stress given each word in a digit string is approximately the same: there are no function words in a digit string, each word is of equal importance, and most words are monosyllabic. Furthermore, the digits do not demonstrate the phonetic richness found in American English.

However, the digits form a suitable vocabulary for studying how each component in our model could be implemented to handle the variations which occur in speech. Although the digit vocabulary does not include all the phones in English, it contains examples of allophonic variations and many low-level phonological effects, such as acoustic gemination and flapping. The phonological effect of acoustic gemination, that is when two similar phonemes merge into one and appear acoustically as one segment, is observed in both consonants and vowels in digit strings. For example, the final /s/ is "six" geminates with the initial /s/ in "seven," such that the two /s/'s appear acoustically as one /s/. The final /s/ in "six" also geminates with the /z/ in "zero." The low-level phonological rule for flapping of /t/'s in an intervocalic position is also illustrated in the digit vocabulary. In the digit strings, the /t/ in "eight" is in intervocalic position when followed by another "eight," and this intervocalic /t/ can be flapped.

Some allophonic variations are observed in digit strings. For example, three realizations of /t/—released, unreleased, and flapped—occur. The released /t/ is usually observed word initially, as in the word "two," or word finally when the word is the last word in a phrase, as in the word "eight" at the end of a sentence. A /t/ in word final but not phrase final position, as in the word "eight" in the middle of a digit string, is frequently unreleased. And a /t/ may be flapped when it occurs in intervocalic position, as in the string "eight eight." At least two realizations of

/v/ are also observed. A strongly fricated /v/ often occurs before a fricative, and a strongly voiced /v/, in which the frication rides on the voicing, occurs primarily in intervocalic positions. In addition, the nasal /n/ occurs in both intervocalic and non-intervocalic position. Thus both robust intervocalic nasals and generally weaker non-intervocalic nasals are illustrated in the digit vocabulary.

The digit vocabulary also illustrates many coarticulation effects, both within words and across word boundaries. The second formant in "two" is raised initially due to the preceding /t/, which is a dental consonant, and it may remain raised if followed by another dental, such as /s/, or it may drop very low in frequency if followed by /w/. Similarly, the formant values of the word final /o^{w}/ in "zero" are strongly affected by the following phonetic environment. Thus the digit vocabulary illustrates a wide variety of phonological variations and coarticulation effects which must be handled by a recognition system.

The digit vocabulary is suitable for demonstrating the utility of the acoustic-phonetic model components. A vocabulary which illustrates the full power of sequential constraints in word candidate reduction leaves little work for the verifier. Consequently, the role of the verifier cannot be studied. For example, the sequential constraints in a small polysyllabic vocabulary may be so strong that most of the word boundaries and most of the words can be specified without ambiguity. Since most of the words are specified, the verifier is not needed except in a few cases. In contrast, the digit vocabulary is primarily monosyllabic and most of the pronunciations of each word contain very few broad phonetic segments. Since some sequential constraints apply, the utility of sequential constraints in recognition can be studied using the digit vocabulary. And since the digit vocabulary does not illustrate the full power of sequential constraints, the role of the verifier can be studied.

In summary, many examples of phonological variation occur in the digit vocabulary, even though it is limited in many ways. Furthermore, the digits form a manageable task for demonstrating how acoustic-phonetic knowledge can be applied

to speech.

5.3 Why a "Preprocessor" Based on Acoustic-Phonetics?

In our recognition model, the broad phonetic classifier and word hypothesizer can be view as a "preprocessor." The purpose of the preprocessor is to rule out unlikely word candidates based upon information that is easy to compute. A preprocessor has been used in other recognition systems as well. For example, an isolated word recognizer developed by Pan et al. (1985) used a preprocessor based on vector quantization (VQ) (see for example, Buzo et al., 1980) to screen word candidates for recognition using dynamic time warping. The primary purpose in their use of a preprocessor was to reduce computational costs while maintaining the performance rate of current DTW systems.

However, an acoustic-phonetic preprocessor has many advantages in addition to possible reduction of computational costs. By using an acoustic-phonetic preprocessor to apply speech constraints to remove poor word candidates, more directed detailed acoustic analysis can be performed. Additionally, an acoustic-phonetic preprocessor segments the speech so the benefits associated with a phonetically-based segment representation can be exploited. These advantages are discussed in more detail in Section 5.4 and Section 5.5.

In our model, an acoustic-phonetic preprocessor screens word candidates based upon broad phonetic information, and the word candidates are evaluated based upon robust information. In contrast, a non-acoustic-phonetically based preprocessor attempts to screen word candidates primarily on detailed spectral information. The purpose of the preprocessor is to rule out unlikely word candidates; attention to fine phonetic differences at an early point in processing is not only unnecessary but is also not as robust. Fine phonetic differences are not as robust because fine

phonetic differences are more sensitive to allophonic variation than differences between a broad phonetic class. This contrast is important because the preprocessor removes word candidates, and the recognizer may not provide for recovery of word candidates which should not have been removed. Thus the decision threshold for removal of a word candidate should be lenient to avoid irrecoverable errors. However, if the information given the preprocessor is not very robust, then the preprocessor can not be efficient in word candidate reduction.

An acoustic-phonetic preprocessor is applicable to continuous speech, as demonstrated in this thesis. In contrast, other preprocessors, such as the VQ-based preprocessor are word based. Therefore, extension of these preprocessors to continuous speech is uncertain because unknown word endpoints must be dealt with.

An acoustic-phonetic preprocessor allows different types of speech information to be incorporated in the identification of broad phonetic classes. Pan et al. found that a VQ preprocessor performed much better when temporal and energy information were also used. However, the information was incorporated into the existing time-frame structure, which is not conducive to using such information.

Most recognition models which do not use an acoustic-phonetic preprocessor recognize an utterance by matching a set of templates to the input signal. Three difficulties that are associated with this recognition method, but are minimized by using an acoustic-phonetic preprocessor are: (1) speaker independence is not easily incorporated, (2) context cannot be explicitly used, and (3) the speech signal is quantized to the template values used in a template representation.

Models which perform spectral matches are inherently speaker-dependent because the spectral representation has not been abstracted to capture the speaker-independent information. DTW and VQ attempt to handle speaker-independence with use of multiple templates. In a network model, multiple paths may be needed to represent different types of speakers. Thus each approach attempts to achieve speaker-independence by capturing variations in sounds, rather than abstracting

the speaker-independent features of sounds, as is possible in an acoustic-phonetic approach. By capturing variations in sounds, measurements made in these approaches are inherently noisy. This is because in addition to information relevant to the speech sounds being identified, information irrelevant to the speech sounds being identified is incorporated into the measurement.

By not using a preprocessor to find a speech motivated recognition unit, these models must use a regularly sampled representation of the input signal. However, a uniformly sampled representation has the undesirable quality that context is difficult, if not impossible, to specify. This is because context is a speech based concept and not a time-frame based concept. In contrast, an acoustic-phonetic preprocessor allows context to be explicitly specified.

Representation of a rising F_2 by a network model with a fixed number of states per phone illustrates the quantization problem. A network model, if it has a sufficient number of states (e.g. Harpy), would try to capture rise in F_2 through a sequence of states. However, the quality of match as the speech signal passes from one state to the next in the sampled representation varies. This variation is not due to variations in the quality of the rising F_2; it is due to quantization in the match. Thus measurements made in network models without a preprocessor do not accurately reflect the events occurring in the speech signal. In contrast, the segments defined by an acoustic preprocessor allow the movement of a formant to be explicitly captured.

In summary, an acoustic-phonetic preprocessor is a valuable part of a recognition system. The preprocessor uses robust information, is applicable to continuous speech tasks, and allows different types of speech knowledge to be easily incorporated into the recognition process. It also provides a basis for performing speaker-independent recognition, allows context to be specified, and allows a more accurate representation of events in the speech signal.

5.4 Why Segments?

An assumption made in this work, and an integral part of the approach is that a sequence of labels may be attached to the speech signal. Furthermore, it was assumed that speech is produced as a sequence of sounds of varying duration which can be represented as a sequence of labels. The sequence of labels, or recognition units, correspond with important phonetic events. As a result, the phonetic recognition units will be irregularly spaced. A phonetic unit can be associated to a (perhaps fuzzy) time in the speech signal or to a (perhaps fuzzy) region of the signal.

This thesis uses phonetic units which were associated with regions of the speech signal, and each region is referred to as a phone segment. This section addresses two issues. First, a segment representation is argued to be superior to a time-frame representation. Second, the advantages of a phonetic segment representation over other segmental representations, such as the diphone and demisyllable, are described.

A segment representation has many advantages over a time-frame representation. For example, a segment which spans a region of speech can be characterized over time. This is an important attribute because many strong cues to speech sounds are distributed across time. For example, systems based on segments or a sequence of phonetic units can explicitly characterize formant motion during the first 30 msec of a vowel to capture transition information. In a frame-by-frame analysis, this information is included, along with other information, only implicitly in the training data.

The use of segments rather than individual spectral analysis frames allows a wider variety of acoustic-phonetic constraints to be exploited. That is, in addition to the ability to characterize transition and relatively stable regions of the speech signal, characterizations over the entire region, such as the maximum, minimum, or average value of a parameter, are available. With the segment formulation, onset

rate can be used to influence the decision on the identity of the whole segment. In contrast, features such as onset rate do not make sense in a frame formulation and would only influence the score in one frame of a spectral distance metric, if at all. Thus a segment framework allows important information to be taken into account explicitly.

As a consequence of the variety of characterizations available with segment-based representations, a system which uses a segment representation can avoid many of the errors produced by systems which simply try to match the spectrum. This is because there is much more information in the speech signal than spectral shape. For example, one striking difference between a strong alveolar fricative and weak dental fricative, given the same context, is the strength of the fricative. Distance metrics such as Itakura's (1975) do not use energy information. Instead, such information must be explicitly incorporated if it is to be used in recognizers based on spectral distance formulations. Researchers are now recognizing the importance of using such information and are devising methods for incorporating such knowledge into existing algorithms, such as vector quantization (Pan et al., 1985; and Bush and Kopec, 1985) and Hidden Markov Modeling (Schwartz et al., 1985). The use of features in an acoustic-phonetic approach provides a unified method for using this knowledge. In an acoustic-phonetic approach, features may be selected based upon human knowledge of what is important, supplemented by statistics to verify that the knowledge has been adequately captured by computer. Many speech motivated segment representations, such as the phone (e.g. Woods et al., 1976), diphone (e.g. Scagliola and Marmi, 1982), syllable (Fujimura, 1975; Mermelstein, 1975), and demisyllable (Rosenburg et al., 1983), have been proposed. We believe that a phonetic representation is better than either the diphone or demisyllable representation. It is more flexible because it can be transformed into a diphone or demisyllable representation. Therefore, all the information which is available from a diphone or demisyllable representation is also available from a phone representa-

tion. Furthermore, many characteristics which are easily computed from a phonetic segment representation are more difficult to extract in a diphone or demisyllable based representation. For example, measurement of duration, such as the duration of a fricative, is straight-forward in a phone representation. In contrast, in a diphone representation, each phone has been split into two parts to form diphones; as a result, information about phone boundaries and phone durations are not easily obtainable.

Acoustic-phonetic knowledge, such as contextual information, can be easily and explicitly expressed using a phone representation. A phone representation is amenable to using information during the relatively stable central portion of the phone and also to using transitional information. This is possible because the phone unit defines regions of the signal which should be stable and also points of transition (the edges of the region). Since acoustic features may be defined over any region of a phone, features characterizing transition regions and features characterizing stable regions can be defined. For example, the feature characterizing the average value of F_1 was computed over the middle 50% of a phone. Since most of the contextual information is contained in the transitions at the beginning and end of a phone, this estimate of F_1 is minimally influenced by context.

By making judgments about sounds based on characteristics over a region of the signal which is generally stationary, such as within a phone, the effect of local variations in the signal can be reduced by techniques such as averaging. Furthermore, estimates of values characterizing the region should be more accurate than the ensemble of single estimates for each point, since the value characterizing a segment is based on examining the data in the region as a whole. Thus the judgments made over a region should be more reliable.

In summary, the use of segments allows a wide variety of acoustic-phonetic constraints to be exploited and a wide variety of characterizations of the speech signal. The flexibility in specifying the importance of information in the signal, available

when the signal is represented by segments, eliminates many errors encountered in spectral matching systems. A phonetic segment representation is superior to other representations because it allows characterization of information available in a diphone or demisyllable representation; furthermore, it allows characterization of other information derived from phonetic units.

In this section we have argued the advantages of a segment representation. However, we should note that we are imposing a segment representation on the speech signal and that a segment representation is a convenience which we use to describe the speech signal.

5.5 Why Use Sequential Constraints in Lexical Access?

Lexical access is the point in the recognition process where information about the speech signal is combined with knowledge about the vocabulary to propose word candidates. Sequential constraints provide a mechanism for removing unlikely word candidates from consideration before the fine discrimination necessary for identification of a phone is performed. The remaining word candidates are relatively similar since each candidate is composed of a string of phones which match the string of broad classes well and therefore match the initial features characterizing a broad class well. As a result, features used initially to identify broad classes are not needed in verification, allowing the verification component to perform more directed analyses.

If sequential constraint application is skipped so that all words are hypothesized at each possible position, then the verifier is burdened with additional phones and words to score. Additional features may be needed since fine as well as gross differences between sounds must be measured, increasing the amount of computation. The contrast between two similar competitors also is reduced; noise may be added

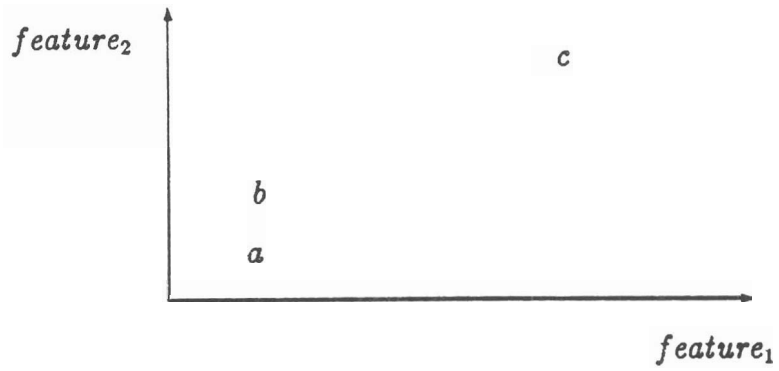


Figure 5.1: Real competitors a and b and outlier c

to the measurements because many other unlikely competitors, such as c in Figure 5.1, are also compared to "real" candidates, a and b, thus adding an "offset" to the scores.

The use of sequential constraints in lexical access thus allows the verifier to focus on the differences between two similar competitors, rather than measuring that a and b are similar relative to c, and also measuring how much better a is to b. As an analogy, if one is trying to measure the peak-to-peak amplitude of an ac signal which is offset by a large dc bias, one would measure on a scale covering only the ac region to get an accurate measurement; one would not include the dc component in the measurement.

In the application of sequential constraints, a risk is associated with ruling out correct word candidates. However, recognition systems based on network models also implicitly use constraints in the search strategy. In continuous speech, the possible paths in a network model based on frame-by-frame analysis can be so large that searching the entire network is intractable. These search strategies generally employ some type of heuristics for pruning, such as the beam search used by the Harpy system. To make a contrast, pruning is applied while searching in a network model, but pruning is applied before creation of a network to be searched in an acoustic-phonetic model. The use of speech constraints, such as sequential

constraints, is advantageous over heuristic search techniques because the risks associated with using speech based constraints can be quantified. However, unless the speech constraints are applied simultaneously, the use of speech constraints cannot be guaranteed to rule out only words which can never achieve a better score.

A network is used to represent the phones in the verification component. The phones in a word are associated with the transitions, and the cost of making a transition is a score which reflects the "goodness" of a hypothesized phone, based upon feature information. The word scores from lexical access may be used in this network representation in several ways. Based directly on the results for using sequential constraints, a pruning threshold can be set to limit the number of word candidates considered. A network can then be constructed from the remaining word candidates. Alternatively, if a heuristic search strategy is used where nodes are expanded as needed, the words candidates could be stored in an ordered queue based on their lexical access score. Nodes could then be expanded using words from the ordered queue. The disadvantage of this method is that the set of phone competitors can change as nodes are expanded. Consequently a phone cannot be scored using discrimination between a phone and its competitors; instead, scores must be assigned based purely on identification information.

In summary, the use of sequential constraints allows the verification component to perform a more detailed and directed evaluation of the phone and word candidates. The risks associated with using speech constraints can be quantified, something which is not as easily achieved using heuristic search algorithms. When constraints are applied to reduce the word candidates at once, competitor phones are known and discrimination between phones can be performed.

5.6 Why a Simple Control Strategy?

The verification component used a simple control strategy: find the best se-

quence of digits by maximizing the score of each phone, where the phone scores are a function of the probability of a hypothesized phone occurring, given the observed feature values. This simple strategy was found to be sufficient for verification of phones in the digit vocabulary.

In a complicated recognition task, a control strategy may be needed which integrates cues in a logical manner and hypothesizes new phones if conflicting cues indicate that more than one phone is present in a region which originally was thought to represent only one phone. An example of a situation in which a new phone should be hypothesized is when clear labial formant transitions are apparent in the left context of a closure, followed by a compact double burst located slightly above F_2 (strong indication of a velar release). Instead of a single stop, two consonants should be hypothesized as occurring over the region: a labial consonant, such as /p$^{\neg}$/, /b$^{\neg}$/, or /v/, followed by a velar stop.

Rather than developing a complex control strategy as outlined above, the simple control strategy was chosen because our understanding of the use and integration of acoustic cues is still very primitive. Furthermore, we are still trying to devise algorithms for effectively capturing the acoustic cues needed by a more complex control strategy.

Thus the emphasis in implementing the verifier was placed on assigning identification scores. Although the earlier example (where two different consonants were hypothesized) is not applicable to the digits, potential conflicts can occur in the digit vocabulary. For example, when acoustic gemination occurs, only one broad phonetic segment is found, but two phones should be hypothesized. Potential conflicts which require additional segments to be hypothesized were avoided by initially hypothesizing all viable candidates. Durational and sequential constraints were used to determine whether a candidate was viable. An identification score for each hypothesized phone was then computed based on the value of well-motivated features and statistics.

The feature values were weighted based upon how well each feature identified the phone in question. Thus a feature which is not relevant to the identification of a phone will be given little weight and minimally affect the score. When a feature indicated the possibility of a hypothesized phone to be contrary to what the other features indicated, this was reflected in the computed score. For example, all features may indicate the possibility of an /s/ to be high, except for one which strongly indicates that an /s/ is very unlikely. The resulting score of the hypothesized /s/ is reduced by an amount which depends on how heavily the conflicting feature is weighed.

The assignment of feature weights assumed that the value of each feature varies over a range, rather than being a conditional value. For example, the value of the feature F_1-Normalized-Position ranged from -1.0 to 1.0. A conditional feature such as whether or not a double burst is observed does not satisfy this assumption. A double burst in the region of F_2 is a strong indicator of a velar release (Cole, et al., 1980). Thus one would like to weight this information conditionally such that when a double burst is observed, the feature capturing this phenomena would be given a very heavy weight. When a double burst is not observed, the feature would be given no weight, and the stop would be identified using other features, such as burst location. The current approach can be extended to handle this type of phenomena.

A sophisticated control strategy would use information about coarticulation in making decisions. The strategy used instead was to select features which are least affected by coarticulation, such as making measurements in the middle 50% of a phone. This strategy was shown to be sufficient for handling coarticulation in the digits.

5.7 Computational Considerations

Recognition systems are currently limited in part by the amount of computa-

tion required, and future recognition systems will have even larger computational requirements if current technologies are simply extended. We believe that an acoustic-phonetic approach has the potential to be used in less restricted tasks because of the types of constraints which are available by using this approach.

Computational requirements can be reduced in an acoustic-phonetic approach by applying speech knowledge to rule out unlikely word candidates. It was shown in this thesis how a particular speech constraint, sequential constraints, can be effectively applied. Even when given input from a front end which produced a non-ideal segmentation, sequential constraints were found to provide a reduction in the number of word hypotheses when the input error characteristics of the front end are known and used. Furthermore, this candidate reduction was achieved with new (broad phonetic) pronunciations of words.

With a phone segment representation, the maximum number of recognition units is limited, and more importantly, the maximum number is independent of the number of words in the vocabulary. Thus the use of a phone representation in an acoustic-phonetic approach gives the approach the potential to be computationally tractable in more unrestricted recognition tasks.

The current implementation used a word spotting approach which is not computationally efficient. This is because the focus of the study was quantification of the constraint provided by knowledge of the lexicon, without interaction from heuristic search techniques, in contrast to the development of a recognition system. However, in implementing a verifier as part of a recognition system, the required computation could be restricted by use of sequential constraints in conjunction with efficient search techniques which expand nodes as needed.

In an acoustic-phonetic approach to continuous speech recognition, the phone endpoints are located before the computation of recognition scores. Thus recognition scores are computed for only one set of endpoints. This approach can be contrasted with template-based approaches which try to find the best set of end-

points by computing a recognition score for each possible endpoint pair. As the vocabulary size increases, the number of possible endpoints which must be considered in a template matching approach increases. Therefore, since recognition scores need not be computed multiple times for different sets of endpoints, the computation required by an acoustic-phonetic approach is more tractable than a template matching approach. We believe that through the use of speech constraints and a segment representation, an acoustic-phonetic approach is a computationally tractable approach for the development of less restrictive speech recognition systems.

Summary

The previous sections have addressed the following issues:

- An acoustic-phonetic approach is appealing intuitively and in its extendability to less restrictive tasks.
- Digits were chosen for the case study because many phonological variations in speech occur in digit strings. Thus the digit task allowed the utility of the model to be demonstrated, yet the constraints on the vocabulary kept the problem manageable.
- An acoustic-phonetic preprocessor is a valuable part of a recognition system because it allows a phone-based segment representation, explicit specification of context, and a more accurate representation of events in the speech signal.
- Segments allow a wider variety of characterizations of the speech signal, eliminating many of the errors encountered in spectral matching systems.
- Sequential constraints reduce the burden on the verifier and allow verification to be more directed and detailed. In addition, the risks associated with application of sequential constraints can be quantified so that they are known.

- A simple control strategy using acoustic-phonetic features based on speech knowledge is sufficient for the digit vocabulary.

- Because of the types of constraints available in an acoustic-phonetic approach, we believe that this approach has the potential to be used in less restricted recognition tasks.

5.8 Contributions of the Thesis

This thesis illustrates that an acoustic-phonetic approach is a viable alternative for building continuous speech recognition systems. This was demonstrated by extending the acoustic-phonetic recognition model for isolated words proposed by Shipman and Zue to continuous speech and by implementing the components in the model for the digit vocabulary case study. Implementation of the components demonstrated how speech knowledge could be used in each component. More importantly, implementation of the components allowed study of the issues involved with handling speech variability when applying speech constraints, a better understanding of how speech constraints can be applied, and a better understanding of how speech constraints and an acoustic-phonetic approach can be used to reduce some of the primary difficulties in developing a speaker-independent, continuous speech recognition system. Although this research used the digit vocabulary as a case study, the observations and issues addressed should be of value in the design of continuous speech recognition systems using other vocabularies.

5.8.1 Extending the Theoretical Model

In this thesis, it was demonstrated that it is feasible to extend the Shipman and Zue model to continuous speech for a limited vocabulary such as the digits. Shipman and Zue's model proposed that sequential constraints applied to *broad class* information from the speech signal provides a significant reduction in the number of

word hypotheses. To extend the model to continuous speech, broad class sequential constraints are used to hypothesize words and also to hypothesize corresponding word boundaries. Performing word hypothesis from a broad phonetic segmentation is in contrast to earlier continuous speech recognition systems, such as HWIM and Hearsay II, which hypothesized words from a phonetic input string. The feasibility study demonstrated that in the digit vocabulary with multiple pronunciations of a word allowed, 66% of the word boundaries could be identified given an error-free input. Furthermore, using path constraints and sequential constraints, an average of 2.8 digits were proposed for every digit in the string. These results show that a significant reduction in word candidates can be achieved using sequential constraints, and, as a result, a recognition model based on sequential constraints can be used for continuous speech.

5.8.2 Contributions from Component Implementation

We explored an acoustic-phonetic model of continuous speech recognition by implementing the model components. By taking the step from a proposed model to implementation of the components of the model, issues important to the application of speech constraints in a recognition system were studied. A method which used front end characteristics for applying sequential constraints to non-ideal data was developed and shown to provide effective candidate reduction. In addition, implementation of the verifier demonstrated the power of an acoustic-phonetic approach which allows a few well-motivated acoustic features to be selected for identification of phones.

Broad Phonetic Classifier

Implementation of the broad phonetic classifier demonstrated several concepts. It illustrated that a set of acoustic features describing robust characteristics of broad classes of speech could be identified. Furthermore, these features could be combined

using a set of production rules to produce a broad phonetic segmentation. Implementation also demonstrated an alternative to earlier segmentation algorithms. Rather than assigning labels to each frame and then grouping the labels to form segments, robust regions, similar to islands of reliability, were identified and then extended outward.

Word Hypothesizer

Implementation of the word hypothesizer illustrated how sequential constraints can be effectively applied to speech. In conjunction, two issues were addressed: (1) Is a segment lattice or a segmentation string a better representation for input to the word hypothesizer? and (2) Can knowledge about front end characteristics and segmentation of speech be combined to produce a score indicating the viability of each word hypothesis?

The variations which occur in real speech dictate that flexibility is required to apply sequential constraints. The segment lattice and segmentation string represent two approaches to handling front end characterizations of speech variations. The segment lattice attempts to handle variations by allowing multiple labels. However, the ambiguity of the lattice makes computation of a meaningful characterization of the broad phonetic classifier questionable. Thus when sequential constraints were applied to a segment lattice, knowledge about the characterization of the broad phonetic classifier was not used. It was found that this method of lexical access did not provide enough flexibility to handle new broad phonetic representations of words, even when a lexicon of alternative pronunciations was used.

In contrast, the insertion, deletion, and substitution characteristics of the broad phonetic classifier can be defined from a segmentation string. A scoring algorithm which computed how well the phonetic pronunciation of a word matched a portion of the segmentation string was developed. The algorithm generally penalized new but similar pronunciations of a word only slightly. The distribution of correct word

scores were observed to cluster near a probability of 1. In contrast, the distribution of the scores of all word hypotheses was very broad. It was shown that these results could be used to set a threshold as an effective way of removing poor word candidates from consideration by the verifier. At the same time, words which had a slightly different broad phonetic representation than the canonical representation derived from the phonetic transcription generally were not ruled out, allowing for new but similar pronunciations of words.

It was observed that path constraints and broad allophonic constraints are not as effective when the number of word hypotheses is large. Thus, in addition to permitting the verifier to be more directed, the use of sequential constraints was found to permit these constraints to be more effective.

Verifier

Implementation of the verification component illustrated that phones are a viable and useful representation. It further illustrated how, in the digit vocabulary, a few (nine) well motivated acoustic features combined with statistical characterizations can be effective in identifying phones. This success was due to the ability to selectively characterize portions of phone segments and also to the ability to variably weight acoustic events which cannot easily be given much importance in a frame-by-frame algorithm. These results thus illustrate the utility of an acoustic-phonetically based decision process. Additionally, it was found that detailed discrimination between competitors provided better identification scores than deriving a score based solely on the characteristics of the phone itself.

5.8.3 Advantages of an Acoustic-Phonetic Approach

Implementation of the model components illustrated how an acoustic-phonetic approach is potentially speaker-independent. In addition, implementation illustrated how an acoustic-phonetic approach can be used to reduce some of the com-

mon difficulties in the development of large vocabulary continuous speech recognition systems, such as coarticulation and speaker-independence.

Speaker-independence

Speaker-independence was examined in each component by comparing the output when the component was tested on new speech by the training speakers and when tested on new speech by new speakers. The output of the broad phonetic classifier was found to be very similar for the two sets of speakers, illustrating that broad phonetic classes can be found independent of speaker.

The lexical access component is affected by speakers in the broad phonetic representation of the speech signal. Since the output of the broad phonetic classifier is relatively speaker-independent, the distribution of the correct word scores, based upon the automatically computed broad phonetic transcription of the speech signal, should also be similar, unless the types of insertions, deletions, and substitutions which occur have changed significantly. It was found that the distribution of correct word scores is essentially the same for training speakers and new speakers on new digit strings. This demonstrates that sequential constraints, when combined with information about front end characteristics, can be used to score word candidates, independent of speaker.

Speaker-independence in the detailed acoustic analysis component was evaluated based on phone and word recognition rates. Comparison of the rank of the correct phone in a word between training speakers and new speakers on new sentences shows no significant degradation. Similarly, comparison of the word error rate between training and new speakers on new speech is essentially the same. Thus, in an acoustic-phonetic approach, one can choose features which key on important information in the speech signal, relatively independent of speaker.

In summary, all three components were found to perform similarly for training speakers and new speakers. Thus an acoustic-phonetic approach shows promise for

speaker-independent recognition.

Coarticulation

Another difficulty in the development of a continuous speech recognition system is coarticulation between words. The use of phone segments as a verification unit allows the system to examine or not examine coarticulation effects. In this thesis, identification of phones by examining regions least affected by coarticulation was studied using the digit vocabulary. It was found that phones can be identified reasonably well using information during the region least affected by coarticulation. In particular, the correct phone was the top candidate at least 87% of the time, and within the top two candidates at least 98% of the time. These results show that use of a phone representation is a powerful method for reducing coarticulation effects to derive a baseline phone identification score. Further work to explicitly exploit transitional information could thus use the best ranking phone candidates as a starting point for verification of possible transitions between phones, based on the acoustic information present.

5.9 Future Work

We believe that the results of this research indicate the viability of an acoustic-phonetic approach to continuous speech recognition and that further studies should be pursued using this model. We also believe that the next step in this area of research is to use a more formalized approach to incorporate speech knowledge into each system component. In this section, we outline suggestions towards a more formalized approach and propose modifications for each of the system components based upon what was learned through implementation of the components.

5.9.1 Broad Phonetic Classifier

Before using the broad phonetic classifier in a recognition system, its performance should be improved. Ideally, the performance should approach a level such that the number of word candidates remaining after application of sequential constraints is comparable to the number of candidates in the ideal word lattice (as was used to evaluate the verifier). The broad phonetic classifier can be improved and extended by using a more formalized approach and by refining the chosen set of broad phonetic labels. Speech knowledge was incorporated into the broad phonetic classifier using heuristics. More formalized methods for defining the set of broad classes used in initial labeling and for identifying segment boundaries and labels are needed. For example, broad phonetic classes could be selected based upon how the phones in a labeled set cluster in a chosen feature space. In addition, a formalized method for identification of segments should still adapt to new utterances; that is, the classification algorithm should use training data as a standard which can be adjusted to utterance characteristics.

The chosen set of broad phonetic classes should be examined in more detail and the least robust classes refined. For example, rather than trying to identify nasals in all contexts, only intervocalic nasal consonants (which are much more robust than word initial or word final nasals in the context of a voiceless consonant) could be included in the sonorant broad class; and non-intervocalic nasals could be included in the vowel broad class. The rules for finding a short voiced obstruent should also be refined. Although a dip in energy is usually observable in higher frequencies, this information was not used; this information could be used by looking for a dip in energy in the higher frequencies and/or by looking at a wider energy band.

The broad phonetic classifier currently segments and labels the speech signal into six broad phonetic classes. By simply extending the approach to larger vocabularies, the number of word candidates will generally increase. Inclusion of several more detailed, but still robust classes, such as labeling vowels as front/back, can reduce

the number of word candidates. In addition, broad classes for glottalization and for aspiration, which frequently occurs at the end of a sentence, especially after a sentence-final /r/, would be helpful since these sounds were not strongly associated with any of the six broad classes.

The boundaries assigned to segments were sometimes offset in time from the hand transcription, resulting in "extra" substitutions. Errors attributable to this offset are more evident in short broad phonetic segments, such as a short voiced obstruent or a word initial nasal. The offset is in part due to arbitrarily dividing the transition regions between the adjacent segments. By reducing the number of offset errors by the broad phonetic classifier, the correct word scores would be more closely concentrated near a probability of 1, and a larger percentage of the word hypotheses could be pruned. Two options for "reducing" the offset error statistics and therefore improving the performance are to: 1) try to find more accurate boundaries in the broad phonetic classifier or 2) compute performance statistics by using a *string matching* algorithm (e.g., Levenshtein distance in Sankoff and Kruskal, 1983) to match the sequence of broad phonetic labels produced by the system with the hand labeled phonetic transcription. In the second option, an assumption is made that the segment boundaries will be adjusted in verification. This option is preferred because an offset in the location of segment boundaries would not be counted as an insertion or deletion. In addition, the location of more accurate segment boundaries is postponed until the identity of the hypothesized phone segments is known.

5.9.2 Lexical Component

In the lexical component, the use of durational, path, and allophonic constraints needs to be further investigated. The number of word candidates remaining after each constraint is applied should be evaluated over a variety of sequential constraint thresholds. Durational constraints were used to specify when a broad phonetic segment could represent only one or only two phones. These constraints were not

used for individual phones because the duration of a phone can vary greatly from utterance to utterance. A reference duration, which could be derived, for example, from knowledge of the number of digits in a digit string, is needed to effectively apply durational constraints at the phone level.

Methods for handling noise in the speech signal, such as lip smacks, need to be incorporated into the application of path constraints. One possibility is to allow possible noise segments to be skipped in a path at a specified cost.

The rules describing allophonic constraints were determined empirically from examination of spectrograms. A more formalized method should be used for defining the rules. For example, the broad phonetic representation (produced by the system) for each pronunciation and context of a word could be statistically tabulated, and these statistics could be used to weight the score assigned to a word. The word scores could then be characterized, as was done in the application of sequential constraints, and a threshold set such that poorly scoring words are removed from further consideration.

5.9.3 Verifier

With a knowledge-based system, more knowledge can always be added. In addition to development of additional detailed acoustic features for verification of phone hypotheses, a method for optimizing the set of features should prove to be valuable in removing features which do not contribute much information.

Allophones of a phoneme were sometimes grouped together in the training statistics. For example, the acoustic realizations of /r/ are different when in prevocalic, postvocalic, and intervocalic position. The current implementation of the verifier primarily used the feature of /r/-Possibility, which was computed over the middle 50% of a phone, to estimate how well a phone is realized as an /r/. This feature favors intervocalic /r/'s, which are most "/r/-like" in the center of a phone, over prevocalic and postvocalic /r/'s. By treating each allophone of /r/ as a different

phone, better results may be obtained.

A weakness in the acoustic-phonetic approach as we implemented it is that the regions of speech used for training and testing were different. Thus, the values obtained in testing may be different. Hence, a method such that training and testing are performed on the same regions of speech should be developed.

A measure of spectral concentration was used to provide rough information about formants, rather than using a formant tracker. When current formant trackers fail, they may make gross errors. The development of a reliable formant tracker or one which indicated the reliability of the computed formant values would be useful in verification. The spectral concentration measure was found to be satisfactory for the limited digit vocabulary; however, a more accurate measure of formant location is needed for other vocabularies.

This research used the digit vocabulary as the basis for exploring how speech constraints can be applied to speech recognition. Pursuing this study for other vocabularies will require development of a more sophisticated control strategy in the verifier, perhaps involving the combination of expert system techniques with multivariate statistics. Coarticulation was largely ignored by defining features over the regions of a phone least affected by coarticulation. With a larger vocabulary, coarticulation will be more important in the identification of phones. One way in which coarticulation effects could be included is by developing a more sophisticated control strategy which can reason about conflicting cues. A control strategy could also be used to handle the varying number of phones in a segment so that normalization by duration or number of segments, both of which have faults, is not necessary.

The verifier was evaluated using incremental simulation. This technique isolated errors due to earlier components from errors due to the verifier, thus allowing the use of acoustic-phonetic features for verification of phones and words to be explored. Before the verifier can effectively use the word lattice produced by the

lexical component, methods need to be developed so that the verifier can handle initial broad class segmentation errors. These methods could include refinement of the broad class boundaries using information about the phonetic transcription of each hypothesized word before verification is performed.

5.9.4 Extensions to Other Tasks

The digit vocabulary formed a well-constrained task in which the utility of speech constraints could be studied in each component of the recognition model. One way to extend this thesis is to modify the knowledge used by the system to include other vocabularies. To extend the system to other vocabularies, the lexicon must be modified to include words from the new vocabulary. In addition, the broad phonetic classifier statistics need to be updated to handle any new phones and phone pair sequences.

A vocabulary could be chosen to illustrate a particular component of the recognition model. For example, a vocabulary composed primarily of dissimilar polysyllabic words could be used to study the performance of lexical access. Ideally, most of the word hypotheses produced by the lexical access component for this vocabulary should be uniquely specified. In contrast, a vocabulary composed of similar words could be chosen to study the use of acoustic features in the verification component.

In summary, this thesis has explored the viability of an acoustic-phonetic recognition model for continuous speech. Using the digits as a case study, the model was shown to be a viable approach to speech recognition which is potentially speaker-independent. We believe that further research should be pursued in order to fully develop this approach.

Appendix A

The Digit Corpus

The digit corpus was composed of 22 seven-digit strings (Corpus A) and 100 random-length digit strings. The random-length digit strings were divided into two subsets, Corpus B and Corpus C. Corpus B was included in the training set so that the system could be trained on random-length digit strings as well as on 7-digit strings. Corpus C was used for evaluating the system components.

The 7-digit strings were defined such that each sequence pair of digits, not including the pair formed by the third and fourth digits, was uniformly represented. The pair formed by the third and fourth digits was not considered in anticipation of people pausing between the third and fourth digit, as when saying a telephone number. However, the speakers were not told to pause, but instead were instructed to say the digit strings as naturally as possible. In addition, the representation of each digit in the string and representation at sentence initial and sentence final position was balanced over the corpus.

The length of the random-length strings was uniformly distributed from one to seven digits. The strings were generated using a random number generator to select the length of the string and the numbers composing the string. The numbers in the string were evaluated for uniform representation of pair sequences. A few strings were edited to insure that each sequence was represented at least once and that

each digit was represented at least once in isolation.

All utterances were orthographically transcribed. In addition, a subset of these utterances was also phonetically transcribed. These transcriptions were manually time-aligned to the speech waveform and with each other using the Spire facility available on the MIT Lisp Machine Workstations (Shipman, 1982; Cyphers, 1985). The position of a segment boundary was determined from observation of the speech spectrogram, the expanded speech waveform, the short-time spectra, and if necessary, by listening to a region of speech.

A few rules were used for transcribing ambiguous cases:

1. Because there usually are no clear boundaries between a vowel and liquid/glide, these boundaries were marked by assigning a fixed proportion of the vocalic region to each label unless the proportion definitely looked incorrect. Non-intervocalic liquid/glides are assigned $\frac{1}{3}$ of the vowel-liquid/glide region. Intervocalic liquid/glides were assigned $\frac{1}{3}$ of the region from the beginning of the preceding vowel to the midpoint of the liquid/glide, plus $\frac{1}{3}$ of the region from the midpoint of the liquid/glide to end of the following vowel.

2. Vowel to vowel boundaries were sometimes difficult to establish when the formants moved smoothly and no glottalization occurred. These boundaries were marked at the middle of the transition between two vowels.

3. When two phones geminate across a word boundary, as in "six seven" or "one nine," the geminate segment was split evenly between the two words unless clear cues to a boundary were evident.

4. When only one release was present in the sequence "eight two," then the closure is assigned to the "eight" and the release is assigned to the "two".

5. Glottalization occurring between two vowels at a word boundary was assigned to the second word.

The utterances were recorded using a Sennheiser noise canceling microphone in a "quiet" room. Corpus A is recordings of four male and three female speakers. Corpus B and Corpus C are recordings of five male and five female speakers. Two of the males and three of the females were the same for the two corpora.

Corpus A			
0315796	1807227	2964898	3674219
4583510	6093882	8240103	9253394
7620085	5471181	6327812	5043023
6861994	7352395	4159706	2869497
8436401	7698316	6077153	5532742
4615931	3214200	2468135	

Corpus B			
43039	90678	88361	56
981357	066	574	69
54	6394829	8846	15
65	019	27581	24
8517	516	733658	7
06	85031	5	38872
307	3		

Corpus C			
9350311	6521325	54434	9
595	89	37	851
1496070	65298	35569	816
5903	432678	91	374
3	547750	0407917	6844
2	6	9005010	61528
57345	15	237	6
2	0	2645	1
4	94	006097	081
4000	52	958399	19120
903	7	50717	1
32	542621	020157	559069
598	260539	686766	40853
305	98959	8762	5
305	0522945	49162	8158054
722	8	5659020	138940
7792	18	7326	541135
780933	13003	66	785753
80	69900		

The utterances in the corpus were divided into four categories:

1. training utterances by training speakers
2. new utterances by training speakers
3. training utterances by new speakers
4. new utterances by new speakers

The numbers followed by a "u" are orthographically but not phonetically transcribed. The utterances in category one form the training set for each of the system components. Thus, the components were trained on random-length digit strings and 7-digit strings spoken by three male and three female speakers.

Subsets of Evaluation Corpus

Speaker	male/female	Corpus Subset		
		A	B	C
jrg	m	1	1	2
mar	m	1	1	2
sch	m	1		
ama	f	1	1	2
cab	f	1	1	2u
chs	f	1	1	2
rhk	m		3	4
wpd	m		3	4
lsp	m		3u	4u
lfl	f		3	4
lsl	f	3	3u	4u

Appendix B

Sample Production Rules

This is an example of a production rule for hypothesizing the phone-like class of strong-fricative-like from acoustic features:

```
(defrule (strong-fric-like1 *initial-rules*)
  (if (and hi-zc
           hi-hfe
           (at-most-one-of vocalic-1 hi-lfe))
      (assert strong-fric-like)))
```

This rule states that if (1) the zero-crossing-rate is high, and (2) the high-frequency-energy is high, and (3) at most one of the indicators of high low-frequency-energy is on, then assert that a strong fricative may be present in the region.

This is an example of a production rule for hypothesizing the phone class of strong-fricative from the phone-like classes:

```
(defrule (strong-fric1 *class-rules*)
  (context-of (((anything)
                ((with-duration > 10 strong-fric-like))
                (anything))
               (if (and (max-greater phfe -65)
                        (max-greater pte -55))
                   (assert strong-fricative)
                   (and (assert strong-fricative)
                        (assert weak-fricative))))))
```

This rule states that if (1) a strong-fric-like segment was hypothesized which has a duration of at least 10 msec, (2) the segment is preceded by anything, and (3) the segment is followed by anything, then a strong-fricative and possibly a weak-fricative is asserted. If the maximum value of high-frequency-energy in the region is greater than -65 dB, and the maximum value of total-energy in the region is greater than -55 dB (the threshold values were determined from a statistical characterization of the phone classes), then only a strong-fricative is asserted; otherwise, both a strong-fricative and a weak-fricative are asserted. Note that durational and contextual constraints can be specified, as well as additional acoustic features.

Appendix C

Insertions and Deletions

Insertions

The following table lists the insertion errors in the output of the broad phonetic classifier which contribute at least 1% of the total number of insertion errors. Note that the errors are reasonable. For example, the two predominant errors are insertion of silence in the labeling of the weak fricatives /f/ and /θ/. As another example, the vowels with offglides form a group in which the offglide portion of the vowel is labeled as a sonorant.

count	broad labels	phone label
25	(SILENCE WEAK-FRIC)	f
25	(SILENCE WEAK-FRIC)	θ
22	(VOWEL SONORANT)	u
19	(SONORANT VOWEL)	ɔ
18	(VOWEL SONORANT)	a^{y}
17	(VOWEL SONORANT)	o^{w}
17	(VOWEL SONORANT)	i^{y}
15	(VOWEL SONORANT)	e^{y}
14	(STRONG-FRIC WEAK-FRIC)	s
13	(SONORANT SILENCE)	n
12	(WEAK-FRIC SILENCE)	f
11	(VOWEL SONORANT)	ü
10	(WEAK-FRIC STRONG-FRIC)	s
8	(WEAK-FRIC SILENCE)	t
8	(VOWEL SONORANT)	n
7	(VOWEL SONORANT)	ɔ
7	(SILENCE STRONG-FRIC)	s
6	(VOWEL SILENCE)	n
6	(VOWEL SONORANT)	r
6	(SILENCE WEAK-FRIC)	v
6	(VOWEL SILENCE)	v
6	(STRONG-FRIC WEAK-FRIC)	t
5	(WEAK-FRIC SONORANT)	θ
5	(SONORANT VOWEL)	r
5	(VOWEL V)	v
5	(SONORANT VOWEL)	n

Deletions

The following table lists the deletion errors in the output of the broad phonetic classifier which contribute at least 1% of the total number of deletion errors. Note that the errors are reasonable. For example, the system was not designed to identify /r/; /r/ was considered to be a vowel. We see that the first four predominant errors are deletion of /r/ when adjacent to a vowel. In addition, the system did not try to locate [k] separately from its closure or the following fricative. Thus the errors of calling [ks] a strong fricative and [k$^{\neg}k$] silence are also reasonable errors.

count	broad label	phone labels
129	VOWEL	r ɚ˞
122	VOWEL	ɪ r
117	VOWEL	r oʷ
90	VOWEL	ɔ r
60	STRONG-FRIC	k s
48	SILENCE	kʰ k
32	VOWEL	ɛ v
26	VOWEL	w ʌ
25	VOWEL	v ə
24	VOWEL	ə n
21	SONORANT	ə n
20	VOWEL	ʌ n
18	VOWEL	ɑ˞ v
18	VOWEL	ʔ e˞
15	VOWEL	ɑ˞ n
14	SONORANT	v ə
14	VOWEL	r ə
13	WEAK-FRIC	□ θ
13	SILENCE	tʰ t
12	VOWEL	ə r
12	STRONG-FRIC	s z
12	WEAK-FRIC	t f
12	SONORANT	n w

Glossary

Some of the terms in this document have different meanings as used by different people. This Glossary is an attempt to clarify the intended meaning of some of these words as used in this thesis.

broad phonetic class: A set of phones which have common acoustic characteristics that can be robustly identified.

broad phonetic level: A description of the speech signal in which the signal is represented as a sequence of segments and each segment is labeled as a broad phonetic class.

constraint: A restriction on the search space. In this thesis, speech knowledge is formulated into constraints which are used to identify poor word hypothesis and rule them out from further consideration

front end: In reference to the acoustic-phonetic recognition model, this is the broad phonetic classifier.

feature/cue: A representation of a region of a parameter which attempts to capture a salient characteristic of the parameter and which may be related to one or more speech sounds.

low level speech knowledge: Characteristics about speech derived from the acoustic signal.

parameter: A set of values directly derived from the speech signal. These values can be used to characterize the speech signal on a sample by sample basis (forming a feature vector) or can be characterized into features. The difference between a parameter and a cue may sometimes be ambiguous.

segment: A region of speech. This region may be associated with a variety of speech units, such as a phone or broad phonetic class.

segmentation string: Representation of the speech signal by a sequence of discrete units, each of which is associated with a label. In the output produced by the broad phonetic classifier, the labels are one of six broad phonetic classes.

sonorant: This term refers to the class of sonorant consonants, in contrast to the distinctive feature used by linguists.

References

Bahl, L.R., A.G. Cole, F. Jelinek., R.L. Mercer, A. Nadas, D. Nahamoo, and M.A. Picheny. "Recognition of Isolated-Word Sentences From a 5000-Word Vocabulary Office Correspondence Task," *Proceedings of the IEEE Internat. Conf. on Acoustics, Speech, and Signal Process.*, pp. 1065-1067, 1983.

Bush, M.A. and G.E. Kopec. "Network-Based Connected Digit Recognition Using Vector Quantization," *Proceedings of the IEEE Internat. Conf. on Acoustics, Speech, and Signal Process.*, pp. 1197-1200, 1985.

Buzo, A., A. Gray, R. Gray, and J. Markel. "Speech Coding Based Upon Vector Quantization," *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. ASSP-28, no. 5, pp. 562-574, 1980.

Chomsky, N., and M. Halle. *The Sound Pattern of English*, Harper and Row: New York, 1968.

Cole, R.A., A.I. Rudnicky, V.W. Zue, and D.R. Reddy. "Speech as Patterns on Paper," in *Perception and Production of Fluent Speech*, Edited by R.A. Cole, Hillsdale, NJ: Lawrence Erlbaum Associates, 1980.

Cole, R.A., M.S. Phillips, S.M. Brill, P. Specker, and A.P. Pilant. "Speaker-independent Recognition of English letters," Paper presented at the 104th meeting of the Acoustical Society of America, Orlando, FL, 1982.

Cole, R.A., R.M. Stern, and M.J. Lasry. "Performing Fine Phonetic Distinctions: Templates vs. Features," in *Invariance and Variability of Speech Processes*, Edited by J. Perkell and D. Klatt, Hillsdale, NJ: Lawrence Erlbaum Associates, forthcoming.

Cook, C.C., and R.M. Schwartz. "Advanced Acoustic Techniques in Automatic Speech Understanding," *Proceedings of the IEEE Internat. Conf. on Acoustics, Speech, and Signal Process.*, pp. 663-666, 1977.

Cyphers, D.S. "Spire: A Speech Research Tool," S.M. thesis, Massachusetts Institute of Technology, 1985.

Doddington, G.R. and T.B. Scholk. "Speech Recognition: Turning Theory to Practice," *IEEE Spectrum*, pp. 26-32, Sept. 1981.

Duda, R.O. and P.E. Hart. *Pattern Classification and Scene Analysis*, New York: John Wiley and Sons, 1973.

Erman, L.D. and V.R. Lesser. "The Hearsay II Speech Understanding System: A Tutorial," in *Trends in Speech Recognition*, Edited by W.A. Lea, Englewood Cliffs: Prentice-Hall, Inc., 1980.

Fant, G. "Distinctive Features and Phonetic Dimensions," *Speech Transmission Laboratory Quarterly Progress and Status Report*, STL-QPSR 2-3, pp. 1-18, 1969.

Fant, G. *Acoustic Theory of Speech Production*, The Hague: Mouton, 1970.

Fujimura, O. "Analysis of Nasal Consonants," *Journal of the Acoustical Society of America*, Vol. 34, No. 12, pp. 1865-1875, 1962.

Fujimura, O. "Syllable as a Unit of Speech Recognition," *IEEE Trans. Acoustics, Speech, and Signal Process.*, vol. ASSP-23, no. 1, pp. 82-87, 1975.

Heffner, R-M.S. *General Phonetics*, Madison: The University of Wisconsin Press, 1950.

House, A.S., and E.P. Neuburg, " Toward Automatic Identification of the Language of an Utterance. I. Preliminary Methodological Considerations," *Journal of the Acoustical Society of America*, Vol. 62, No. 3, pp. 708-713, 1977.

Hyde, S.R. "Automatic Speech Recognition: A Critical Survey and Discussion of the Literature," in *Human communication: A Unified View*, Edited by E.E. David, Jr. and P.B. Denes, New York: McGraw Hill, pp. 399-438, 1972.

Itakura, F. "Minimum Prediction Residual Principle Applied to Speech Recognition," *IEEE Trans. Acoustics, Speech, and Signal Process.*, vol. ASSP-23, pp. 67-72, 1975.

Jakobson, R., C.G.M. Fant, and M. Halle. *Preliminaries to Speech Analysis: The Distinctive Features and their Correlates*, Cambridge, Mass.: The MIT Press, 1952.

Jelinek, F., L.R. Bahl, and R.L. Mercer. "Design of a Linguistic Statistical Decoder for the Recognition of Continuous Speech" *IEEE Trans. Infor. Theory*, vol. IT-21, pp. 250-256, May 1975.

Jelinek, F. "Continuous Speech Recognition by Statistical Methods," *Proceedings of the IEEE*, vol. 64, No. 4, pp. 532-556, 1976.

Jelinek, F., R.L. Mercer, and L.R. Bahl. "Continuous Speech Recognition: Statistical Methods," unpublished paper, 1980.

Jelinek, F. "Self-Organized Continuous Speech Recognition," *Proceedings of the NATO Advanced Summer Inst. Auto. Speech Analysis and Recognition*, Bonas, France, 1981.

Kaplan, G. "Words into action:I," *IEEE Spectrum*, pp. 22-26, June 1980.

Kato, Y. "Words into action III: a commercial system," *IEEE Spectrum*, p. 29, June 1980.

Klatt, D.H. "Review of the ARPA Speech Understanding Project, *Journal of the Acoustical Society of America*, Vol. 62, No. 6, pp. 1345-1366, 1977.

Lea, W.A., ed. *Trends in Speech Recognition*, Prentice Hall Inc., Englewood Cliffs, New Jersey, 1980.

Lesser, V.R., R.D. Fennell, L.D. Erman, and D.R. Reddy. "Organization of the Hearsay II Speech Understanding System," *IEEE Trans. Acoustics, Speech, and Signal Process.*, vol. ASSP-23, pp.11-24, 1975.

Lowerre, B. "Dynamic Speaker Adaptation in the Harpy Speech Recognition System," *Proceedings of the IEEE Internat. Conf. on Acoustics, Speech, and Signal Process.*, pp. 788-790, 1977.

Lowerre, B., and D.R. Reddy. "The Harpy Speech Understanding System," in *Trends in Speech Recognition*, Edited by W.A. Lea, Englewood Cliffs: Prentice-Hall, Inc., 1980.

Mermelstein, P. "A Phonetic-Context Controlled Strategy for Segmentation and Phonetic Labeling of Speech," it IEEE Trans. Acoustics, Speech, and Signal Process., vol. ASSP-23, pp. 79-82, 1975.

Myers, C.S., and L.R. Rabiner. "A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition," *IEEE Trans. Acoustics, Speech, and Signal Process.*, vol. ASSP-29, pp. 284-297, 1981a.

Myers, C.S., and L.R. Rabiner. "Connected Digit Recognition Using a Level-Building DTW Algorithm," *IEEE Trans. Acoustics, Speech, and Signal Process.*, vol. ASSP-29, pp. 351-363, 1981b.

Oppenheim, A.V., and R.W. Schafer, "Homomorphic Analysis of Speech," *IEEE Trans. on Audio and Electroacoustics*, Vol. AU-16, No. 2, pp. 221-226, 1968.

Otten, K.W. "Approaches to the Machine Recognition of Conversational Speech," in *Advances in Computers*, Edited by F.L. Alt, M. Ribinoff, and M.C. Yovits, New York: Academic Press, Vol. 11, pp. 127-163, 1971.

Pan, K.C., F.K. Soong, L.R. Rabiner, and A.F. Bergh. "An Efficient Vector-Quantization Preprocessor for Speaker Independent Isolated Word Recognition," *Proceedings of the IEEE Internat. Conf. on Acoustics, Speech, and Signal Process.*, pp. 874-877, 1985.

Pavlidis, T. and S.L. Horowitz. "Segmentation of Plane Curves," *IEEE Trans. on Computers*, vol. C-23, no. 8, pp. 860-870, 1974.

Peterson, G.E. and H.L. Barney. "Control Methods Used in a Study of the Vowels," *Journal of the Acoustical Society of America*, 24(2) pp. 175-184, 1952.

Rabiner, L. "On Creating Reference Templates for Speaker-Independent Recognition of Isolated Words," *IEEE Trans. Acoustics, Speech, and Signal Process.*, vol. ASSP-26, no. 1, pp. 34-42, 1978.

Rosenburg, A.E., L.R. Rabiner, J.G. Wilpon, and D. Kahn. "Demisyllable-Based Isolated Word Recognition System" *IEEE Trans. Acoustics, Speech, and Signal Process.*, vol. ASSP-31, no. 3, pp. 713-726, 1983.

Sakoe, H. "Two-Level DP-Matching–A Dynamic Programming-Based Pattern Matching Algorithm for Connected Word Recognition," *IEEE Trans. Acoustics, Speech, and Signal Process.*, vol. ASSP-27, pp. 588-595, 1979.

Sankoff, D. and J. Kruskal (eds.), *Time Warps, Strong Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, Reading, Mass.: Addison-Wesley, 1983.

Scagliola, C. and L. Marmi. "Continuous Speech Recognition via Diphone Spotting. A Preliminary Implementation," *Proceedings of IEEE Internat. Conf. on Acoustics, Speech, and Signal Process.*, pp. 2008-2011, 1982.

Schwartz, R., Y. Chow, O. Kimball, S. Roucos, M. Krasner, and J. Makhoul. "Context-Dependent Modeling for Acoustic-Phonetic Recognition of Continuous Speech," *Proceedings of IEEE Internat. Conf. on Acoustics, Speech, and Signal Process.*, pp. 1205-1208, 1985.

Schwartz, R.M. and V.W. Zue. "Acoustic-Phonetic Recognition in BBN SPEECHLIS," *Proceedings of IEEE Internat. Conf. on Acoustics, Speech, and Signal Process.*, pp. 21-24, 1976.

Shipman, D.S. and V.W. Zue. "Properties of Large Lexicons: Implications for Advanced Isolated Word Recognition Systems," *Proceedings of IEEE Internat. Conf. on Acoustics, Speech, and Signal Process.*, pp. 546-549, 1982.

Shipman, D. "Development of Speech Research Software on the MIT Lisp machine," Paper presented at the 103rd meeting of the Acoustical Society of America, Chicago, Il., April 1982.

Sigurd, B. "Phonotactic Aspects of the Linguistic Expression," in *Manual of Phonetics*, 2nd ed., Edited by B. Malmber, Amsterdam: North Holland Publishing Co., 1970.

Winston, P.H. *Artificial Intelligence*, 2nd ed., Reading, MA: Addison-Wesley, 1984.

Wolf, J.J. and W.A. Woods. "The HWIM Speech Understanding System," in *Trends in Speech Recognition*, Edited by W.A. Lea, Englewood Cliffs: Prentice-Hall, Inc., 1980.

Woods, W., M. Bates, G. Brown, B. Bruce, C. Cook, J. Klovstad, J. Makhoul, B. Nash-Webber, R. Schwartz, J. Wolf, and V. Zue. "Speech Understanding Systems: Final Technical Progress Report," Bolt Beranek and Newman, Inc, Report No. 3438, Vol. II, Cambridge, Mass., 1976.

Woods, W.A. "Optimal Search Strategies for Speech Understanding Control," in *Readings in Artificial Intelligence*, Edited by B.L. Webber and N.J. Nilsson, Palo Alto, Calif.: Tioga Publishing Company, 1981.