The Use of Distinctive Features for Automatic Speech Recognition

by

Helen Mei-Ling Meng
S.B., Massachusetts Institute of Technology
(1989)

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

Master of Science

at the

Massachusetts Institute of Technology
May, 1991

Signature of Author .....
Department of Electrical Engineering and Computer Science
May 10, 1991

Certified by ....
Victor W. Zue
Principal Research Scientist,
Department of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by ....
Arthur C. Smith
Chair, Department Committee on Graduate Students

The Use of Distinctive Features for Automatic Speech Recognition

by

Helen Mei-Ling Meng

Submitted to the Department of Electrical Engineering and Computer Science
in May, 1991 in partial fulfillment of the requirements for the degree of
Master of Science

Abstract

One of the most critical and yet unsolved problems in phonetic recognition is the transformation of the continuous speech signal to a discrete representation for accessing words in the lexicon. In order to find an efficient description of speech for recognition tasks, our research investigates the use of distinctive features. Distinctive features are a small set of linguistic units which have the potential advantage of enabling us to describe contextual and coarticulatory variations in speech more parsimoniously and thus make more effective use of available training data.

To access the usefulness of distinctive features, we focus our inquiry on three questions. First, is there a particular spectral representation that will yield superior performance over others? Second, how would the extraction and use of acoustic attributes affect classification performance when compared to the direct use of the spectral representation? Finally, are there performance advantages in introducing an intermediate linguistic representation between the signal and the lexicon?

Our investigation lies within the scope of classifying American English vowels using a multi-layer perceptron classifier with a single hidden layer. Vowel tokens were extracted from the TIMIT corpus. To answer the first question, several spectral representations were compared. The combination of the outputs from Seneff's Auditory Model outperformed all other representations with both clean and noisy conditions, yielding top-choice accuracies of 66% and 54% respectively. To answer the next two questions, classification experiments were conducted under six different conditions, which resulted from systematically varying three condition variables. These variables specify whether acoustic attributes were extracted, whether an intermediate feature-based representation was introduced, and how the feature values were combined. Potential computational and descriptive advantages were shown for acoustic attributes and features, respectively.

Thesis Supervisor: Victor W. Zue
Title: Principal Research Scientist

To my family

Acknowledgements

My deepest gratitude goes to my thesis supervisor, Victor Zue, for giving me the opportunity to learn from his many areas of expertise, and providing me with an excellent research environment. I thank him for his invaluable guidance and stimulating advice, as well as his constant encouragement, support and humour, all of which is most crucial to the successful completion of this thesis.

My whole-hearted thanks also goes to the many high-caliber individuals of the Spoken Language Systems Group for their friendship and assistance. Their suggestions and discussions have benefited this work greatly. In particular, I would like to thank:

Hong Leung, for encouraging me to join this group in the first place, for generously offering his expertise in ANN in both ideas and programs,

Jim Glass, for his help with APACHIE,

Mike Phillips, for his help with SAILS,

Stephanie Seneff, for reading earlier drafts of this thesis, and for her comments and suggestions,

Nancy Daly, Jim Glass, Caroline Huang, Jeff Marcus, John Pitrelli and Stephanie Seneff for teaching me how to read spectrograms,

Dave Goodine and Joe Polifroni, for keeping the Lisp Machines running smoothly,

Rob Kassel for creating a marvellous environment for making documents, and David Whitney, for his help in ensuring that the laser printers work,

Vicky Palay, for her assistance in many ways.

I would also like to thank Professor K. Stevens for his interest in my work, and his many helpful insights and discussions.

Finally, I would like to thank my parents and my brothers for their love, for their endless encouragement and unfailing support throughout my years of study at M.I.T., for their comforting words during times of hardships, and most important of all, for instilling in me a love for learning.

This research is supported by DARPA under Contracts N00014-82-K-0727 and N00014-89-1332, monitored through the Office of Naval Research.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem Statement and Motivation

Human-machine interaction via speech has always been a dream and a goal for many people, since speech is regarded as the most natural and efficient means of communication for humans. However, despite active research in the field of automatic speech recognition over the decades, the performance of current technology in restricted domains such as limited vocabulary, isolated word and speaker dependent tasks still falls below human capabilities. One of the most critical and yet unsolved problems is the transformation of the continuous speech signal into a discrete representation for accessing words in the lexicon. To tackle this problem of speech decoding, it is important for us to understand how speech can be represented.

Languages can be described in terms of a small set of abstract linguistic units called *phonemes* [7]. A phoneme is the basic contrastive unit in the phonology of a language. Several phonemes concatenated together constitute a word. Therefore, words with different phoneme sequences are differentiated in a language. For example, the word "hat" consists of the phonemes /h/, /æ/ and /t/ and changing the middle phoneme to /i/ results in the word "heat". Another example is the word "bow" which consists of the phonemes /b/ and /ɑ^{w}/, but inserting the phoneme /r/ in between results in the word "brow". Each phoneme is produced by a unique articulatory gesture, and based on sim-

ilarities and differences in these articulatory characteristics, phonemes can be grouped into classes and sub-classes [28]. In particular, the American English language has 40 phonemes, which can be grouped into *vowels* and *consonants.* The vowels can be further divided into *monophthongs* and *diphthongs*, whereas the consonants can be categorized into *semi-vowels, nasals, stops, fricatives* and *affricates.*

The acoustic signal produced when a phoneme is pronounced is subjected to a wide range of variabilities, since the articulatory movements are continuous and can vary in uncountably many ways. There are contextual and coarticulatory effects, where the realization of a phoneme is dependent upon the identities of the neighboring phonemes. For example, the phoneme /s/ in "gas" is often *palatalized* to become /š/ in "gas shortage". Due to the continuous movement of the articulatory organs under inertia, sharp transitions from one phoneme to another may not always be produced. The direction of these phonological effects is not always consistent, as can be reflected by the absence of *palatization* of /s/ in a /š/ context in the example "tuna fish sandwich". To a certain extent, these phonological effects are imposed by the speaker. There are variations across speakers, as well as variations within the same speaker. Factors such as dialect, vocal tract shape, speaking style, speaking rate, etc., all play a part in modifying the resultant acoustic outcome of a phoneme. In addition, there are environmental factors due to recording equipment and noise. Therefore, the task of classifying a given acoustic segment as a phoneme is immensely complicated due to the wide range of variabilities mentioned above, and classification accuracy will be forseeably low, even though we may reference a large number of examples of each phoneme in the training data.

In order to account for the physical sound produced more accurately, the *phone* has been used as a descriptive unit. The term *allophone* is used to describe a class of phones which are variants of the same phoneme [25]. For instance, the allophone of /t/ in "butter" is realized as a flap, which involves

a quick movement of the tongue tip to and away from the roof of the mouth. Phones can account for sounds in the speech signal very precisely, but there is no objective limit to the number of phones necessary to describe the speech signal. In other words, the coverage of any arbitrarily selected inventory of phones is not complete. This poses some limitations to the use of phones in speech recognition. A large inventory of phones is necessary for reasonable acoustic coverage, which naturally demands a vast amount of training data. Furthermore, should a new phone be discovered and added to the inventory, additional training data and acoustic models will be required. Consequently, systems which utilize the phone as a descriptive unit of speech may not achieve very high adaptability.

At this point, we may perhaps generalize the characteristics of a desirable inventory of phonological descriptive units. The inventory should be small and capable of describing a broad range of sounds. This demands efficiency in capturing phonemic similarities and contrasts due to coarticulation, thereby minimizing the amount of redundancy in the description. The description should also be robust towards environmental variations such as noise. In addition, it should be salient in the acoustic signal for easy identification. A potentially better alternative to the use of phones is offered by *distinctive features*, which will be described in detail in the following section.

1.2 Distinctive Features

The concept of distinctive features is very powerful for analyzing speech. Linguists generally believe that phonemes can be represented by a small set of basic linguistic units - distinctive features [2]. A *feature* is a minimal unit which distinguishes a pair of maximally close phonemes. For example, /**b**/ and /**p**/ are distinguished by the feature [VOICE]. The description corresponds directly to contextual variability and coarticulatory phenomena. For instance, the vowel in "dwell" is probably underlyingly an /**ɛ**/ with an exceptionally

low second formant, since it is influenced by the feature [ROUND] from the left context, which refers to the rounding of the lips in pronouncing /w/, and the feature [LATERAL] from the right context, which associates with the raising of the tongue towards the palatal midline during the articulation of /l/. The complete set of distinctive features can thus describe all phonemically relevant differences occuring with all possible contrasting phoneme pairs. Phonemes sharing features in common form natural classes, e.g. *nasals*, and sounds are more often confused in relation to the number of features they share. It is believed that around 15 to 20 distinctive features are sufficient to account for phonemes in all languages of the world.

Distinctive features are linguistically motivated, and manifest themselves as their corresponding acoustic correlates in the speech signal. Phonological and phonetic research conducted over the past three decades has resulted in a wealth of information, albeit incomplete, on the acoustic correlates of distinctive features. Some of the findings and ideas are presented in the following:

Fant's 'segmental theory' of speech [6] regards connected speech as segments - the temporal contrasts are described by *manner features*, and continuous variations within the segments or across segment boundaries are described by *place features*. A manner feature correlates with the speech wave through its production characteristics, for example, the feature [VOICE] is characterized by the vocal cord vibrations modulating an air stream, which causes the speech wave to have quasi-periodic fine structure in frequency and time. A place feature correlates with the speech wave through its articulatory characteristics. For example, the feature [ROUND] is realized by protruding the lips and drawing them relatively close, resulting in the lowering of the first three formants (especially F2 in most cases) in the speech signal. Therefore, it can be seen that the acoustic correlates of distinctive features tend to be quite localized in the speech signal. Features can co-occur and reinforce other features, and in some cases, certain features provide markers that indicate regions where properties associated with other features are evident in the sound. For

example, evidence for vocal-fold vibration associated with the feature [VOICE] usually occurs in the vicinity of changes in the [+CONSONANTAL] property.

Stevens defined the acoustic correlates of distinctive features using a different approach [35]. He observed many examples of a non-monotonic or "sigmoidal" relation between acoustic and articulatory parameters as schematized in Figure 1.1. As the articulatory parameter is varied gradually, there are ranges where the acoustic parameter is relatively invariant, but as the articulatory parameter moves through the rapid transition region of the sigmoid, the acoustic parameter undergoes a qualitative change. Similar phenomena have been observed between auditory and acoustic parameters. He suggested that these "quantal" relations play a principal role in shaping the inventory of articulatory states and their acoustic consequences that are used to signal *distinctions* in language. The acoustic attributes that occur in the plateau-like regions of the relations are the acoustic correlates of the distinctive features. In his examples, these acoustic correlates should be described in relational terms. This may make distinctive features a purer representation of speech, because relational parameters are more likely to be independent of vocal-tract size, speaking rate, and phonetic contexts than absolute parameters such as frequencies of spectral components. Therefore, Stevens suggested that an utterance in speech may have an underlying representation in terms of distinctive features, possibly expressed as a hierarchy of matrices.

Despite all the information we have about the acoustic correlates of distinctive features, many questions still exist. The hierarchical structure of distinctive features is not completely known. The acoustic correlates of some features have not been fully understood and characterized. It is also uncertain whether the features should be assigned binary values, and how much orthogonality exists between different features. But nevertheless, there are reasons to believe that the concept of distinctive features is potentially very useful for automatic speech recognition. The compact inventory of features enables us to make more effective use of training data. The descriptive power

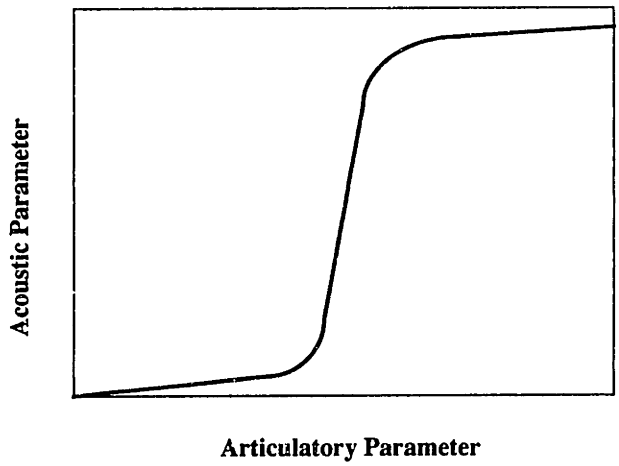

Figure 1.1: Schematization of the quantal relation between an acoustic and an articulatory parameter

of features allow us to account for contextual influence more parsimoniously. For example, the vowel /u/ occuring in an alveolar context is often fronted to become /ü/. So instead of carrying two separate acoustic models for these two vowels respectively, we may perhaps simply note that they share most features except for [BACK], as a result of context, and therefore the feature [BACK] is distinctive. In some cases, coarticulatory effects provide redundant sources of information about the adjacent phonemes, and this may contribute to sustaining high recognition performance. Furthermore, distinctive features can serve as a powerful data reduction and refinement scheme that can be used to save computation, since it may be possible to describe speech by extracting the acoustic correlates of distinctive features, instead of using the entire spectral representation.

1.3 Decoding Strategies of the Speech Signal

Our next step is to explore the use of distinctive features in decoding the speech signal, where an acoustic representation is mapped to the lexicon. Specifically, in this thesis we focus on phonetic classification. One possible method is to in-

troduce an intermediate representation of linguistic features between the signal and the lexicon. This approach, as illustrated in Figure 1.2, clearly offers more flexibility than the direct classification of phonemes from the signal. However, it is not clear whether a set of acoustic attributes is required to bridge the gap between the acoustic representation and the phonological representation. Distinctive features manifest themselves as their acoustic correlates in the speech signal, and phonetic contrasts are therefore inherent in the signal. We cannot as yet clearly characterize acoustic correlates of the various distinctive features, but it is very likely that each feature relates to a region in the acoustic space, and there is a great deal of overlap among such regions. In other words, the acoustic correlates may exhibit varying degrees of prominence in the acoustic signal, and some acoustic representations may be more revealing than others with regard to the underlying features. Moreover, in the process of mapping the acoustic representation to the intermediate feature representation, it may be constructive to extract some acoustic attributes which enhance feature characteristics. Alternatively, since these acoustic attributes are based on the distinctive features, the phonological representation may be bypassed entirely. Amongst these several approaches to decode the speech signal, which have all been included in Figure 1.2, it is not certain which would be the best strategy. Therefore, the objective of this thesis is to assess the usefulness of distinctive features for phonetic classification, and compare the different methods of introducing them as an intermediate representation in our classification framework.

1.4 Thesis Overview

In this thesis, we attempt to address issues related to the use of distinctive features for phonetic classification. More formally, we ask three questions. First, is there a particular spectral representation that is preferred over others? Second, should we use the spectral representation directly for phoneme/feature

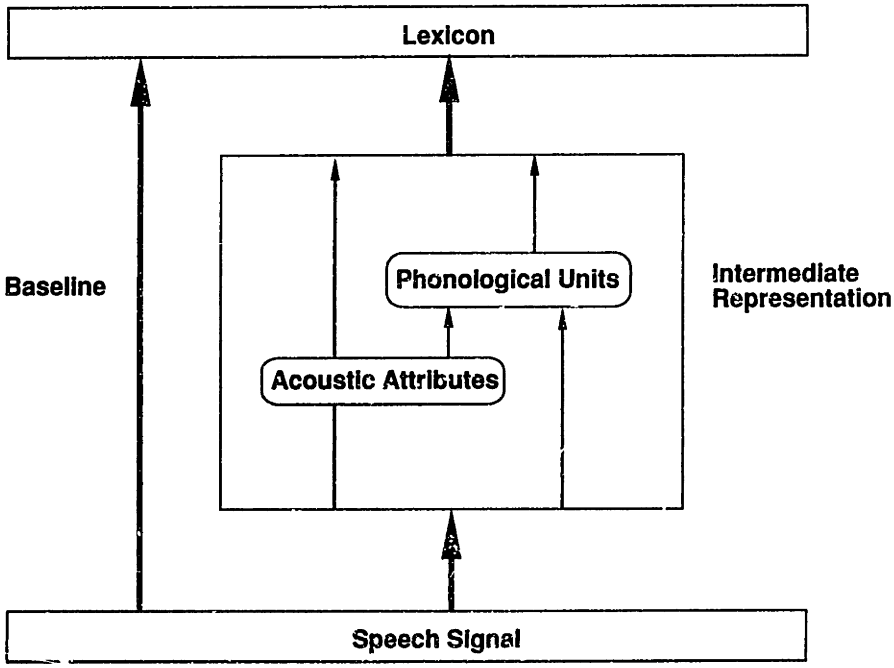

Figure 1.2: Using an Intermediate Representation in the Process of Decoding the Speech Signal

classification, or should we instead extract and use acoustic attributes? Finally, does the introduction of an intermediate feature-based representation between the signal and the lexicon offer performance advantages?

To provide an answer to the first question, we conduct a set of phoneme classification experiments using a variety of input representations. This is described in Chapter 2. We may infer that the representation which gives the best performance should also be the most suitable for use in defining and quantifying acoustic attributes corresponding to the distinctive features.

Then in Chapter 3 we proceed to evaluate the different strategies for decoding the speech signal. Our experimental paradigm includes the baseline approach where the acoustic signal is directly used for phonetic classification. Another approach involves extracting acoustic attributes from the signal before phonetic classification is done. A third approach introduces an intermediate phonological representation between the signal and the lexicon, and the final approach includes both attribute extraction and an intermediate representation.

Following this, Chapter 4 compares several acoustic representations on the basis of their ability to perform acoustic segmentation. Phonemes are the smallest unit which are concatenated to form speech. A sequence of phonemes may constitute a small structure such as a syllable or a word, or a large structure like a phrase or a sentence. This *sequential phonological description* is manifested as a *segmental acoustic description*, and there are clearly overlaps from segment to segment. In this respect, a descriptive acoustic representation should preserve acoustic regularities within a segment which lies between acoustic landmarks, as well as transitional acoustic behavior which occurs across segment boundaries.

The final chapter presents a summary of this thesis as well as possible extensions for future work.

Chapter 2

Selecting an Acoustic Representation

In the selection of an optimal signal representation for an automatic speech recognition system, it is important to bear in mind that the parametric representation should preserve all the relevant aspects of the speech signal for the recognition task in hand and eliminate the irrelevant details. The representation should also be compact, for the sake of computational economy.

Historically, short time spectral representations used as input to a recognizer have included those based on the Discrete Fourier Transform, as well as those based on the all-pole modelling of speech (Linear Predictive Analysis) [28]. Since it is believed that speech is optimized through the evolution of language for the characteristics of human hearing, and there is physiological and psychoacoustical evidence that the ear performs spectral analysis on the speech signal, researchers have built front ends which emulate natural auditory processing [1,3,13]. In some cases, such auditory models have helped to improve recognition performance. There are also the *mel-frequency* representations [16], which is an engineering approximation of the ear's critical band filtering, and has recently gained popularity in the speech recognition community.

2.1 Previous Work with Comparison of Parametric Representations

Several experiments on comparing signal representations have been reported in the past. Mermelstein and Davis [16] compared five representations, namely the mel-frequency cepstral coefficients (MFCC), the linear frequency cepstral coefficients, the linear prediction cepstrum, the linear prediction spectrum, and the reflection coefficients. On the task of recognizing monosyllabic words spoken continuously by two speakers, they found that a set of 10 MFCC resulted in the best performance, suggesting that the mel-frequency cepstra possess significant advantages over the other representations.

Hunt and Lefebvre [14] compared the performance of their psychoacoustically-motivated auditory model with that of a 20-channel mel-cepstrum. The first eight discriminant functions obtained by applying linear discriminant analysis on the two auditory model outputs were compared with 8 unweighted MFCC (C_1 to C_8). Experiments conducted include speaker-dependent and independent conditions, connected and quasi-isolated word recognition, as well as noisy and spectrally tilted speech. The auditory model gave the highest performance under all conditions, and is least affected by changes in loudness, interfering noise and spectral shaping distortions.

Later, Hunt and Lefebvre [15] conducted another comparison with the auditory model output, the mel-scale cepstrum with various weighing schemes, cepstrum coefficients augmented by the δ-cepstrum coefficients, and the IMELDA representation which combined between-class covariance information with within-class covariance information of the mel-scale filter bank outputs to generate a set of linear discriminant functions. The tests conducted were similar to those in the previous comparison. The IMELDA outperformed all other representations.

In summary, these studies generally show that the choice of parametric representations is very important to recognition performance, and auditory-based

representations generally yield better performance than more conventional representations. In the comparison of the psychoacoustically-motivated auditory model with MFCC, however, different methods of analysis led to different results. Therefore, it will be interesting to compare outputs of an auditory model with the computationally simpler mel-based representation when the experimental conditions are more carefully controlled.

2.2 Overview of the Comparison Experiments

This chapter describes a comparative study of six acoustic representations on the task of vowel classification using an artificial neural net (ANN) classifier. Three of the representations are obtained from the auditory model proposed by Seneff [31,30]. Two representations are based on mel-frequency, and the remaining one is based on the conventional Fourier transform. Attention is focused upon the relative classification performance of the signal representations, the effect of increasing training data on the robustness of the results, and the tolerance of the different representations to additive white noise.

To strive towards a fair comparison of the various signal representations, we restricted the ANN classifier to have the same architecture throughout the experiments. All input feature vectors were measured at the same points in the speech signal, and the dimensionalities of the input vectors were all identical.

2.3 Signal Processing

The speech signal is sampled at 16 kHz and a spectral vector is computed once every 5 ms. Three feature vectors, representing the average spectra for the initial, middle, and final third of every vowel token, are determined for each representation. These vectors attempt to crudely capture the dynamic characteristics of vowel articulation. All the acoustic representations result in a 40-dimensional feature vector covering a frequency range of slightly over 6 kHz.

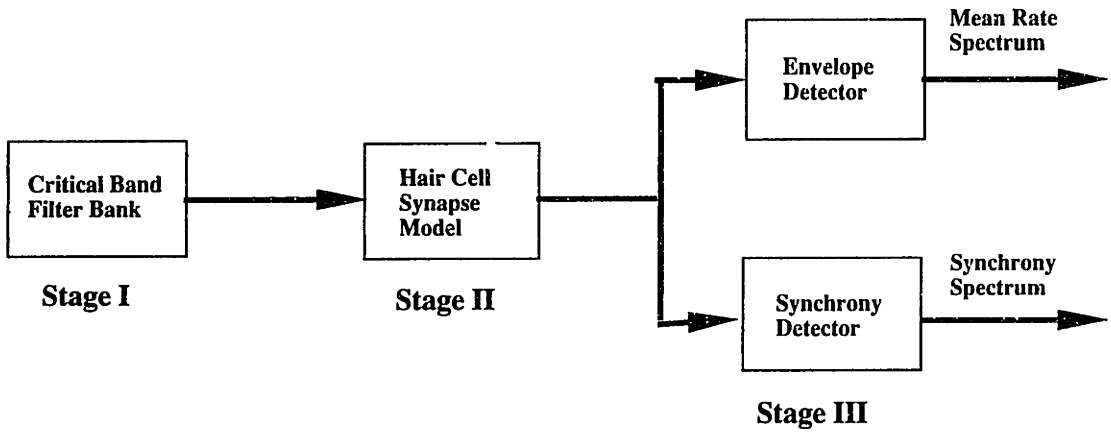

Figure 2.1: Block diagram of Seneff's auditory model

2.3.1 Seneff's Auditory Model

Seneff's Auditory Model (SAM) has three stages [30], as illustrated in Figure 2.1. Stage I consists of a bank of 40 critical band filters, spaced linearly on a Bark frequency scale. The center frequencies of these filters range from 130 to 6400 Hz, as shown in Figure 2.2. The outputs of this stage, the critical band envelopes, are fed into Stage II, which models the transformation from the basilar membrane vibration to the the auditory-nerve fiber responses. This part of the model incorporates non-linearities such as dynamic range compression, half-wave rectification, short-term and rapid adaptation, and forward masking. The output of this stage represents a probability of firing along the auditory-nerve. This will be processed by the envelope detector in Stage III to become the mean probability of firing along the auditory nerve, called the *mean rate response.* The other module, the synchrony detector, determines the synchronous response of each filter by measuring the extent of dominance of information at the filter's characteristic frequency. This output is therefore called the *synchronous response.* Both the mean rate and the synchronous responses result in a 40-dimensional feature vector.

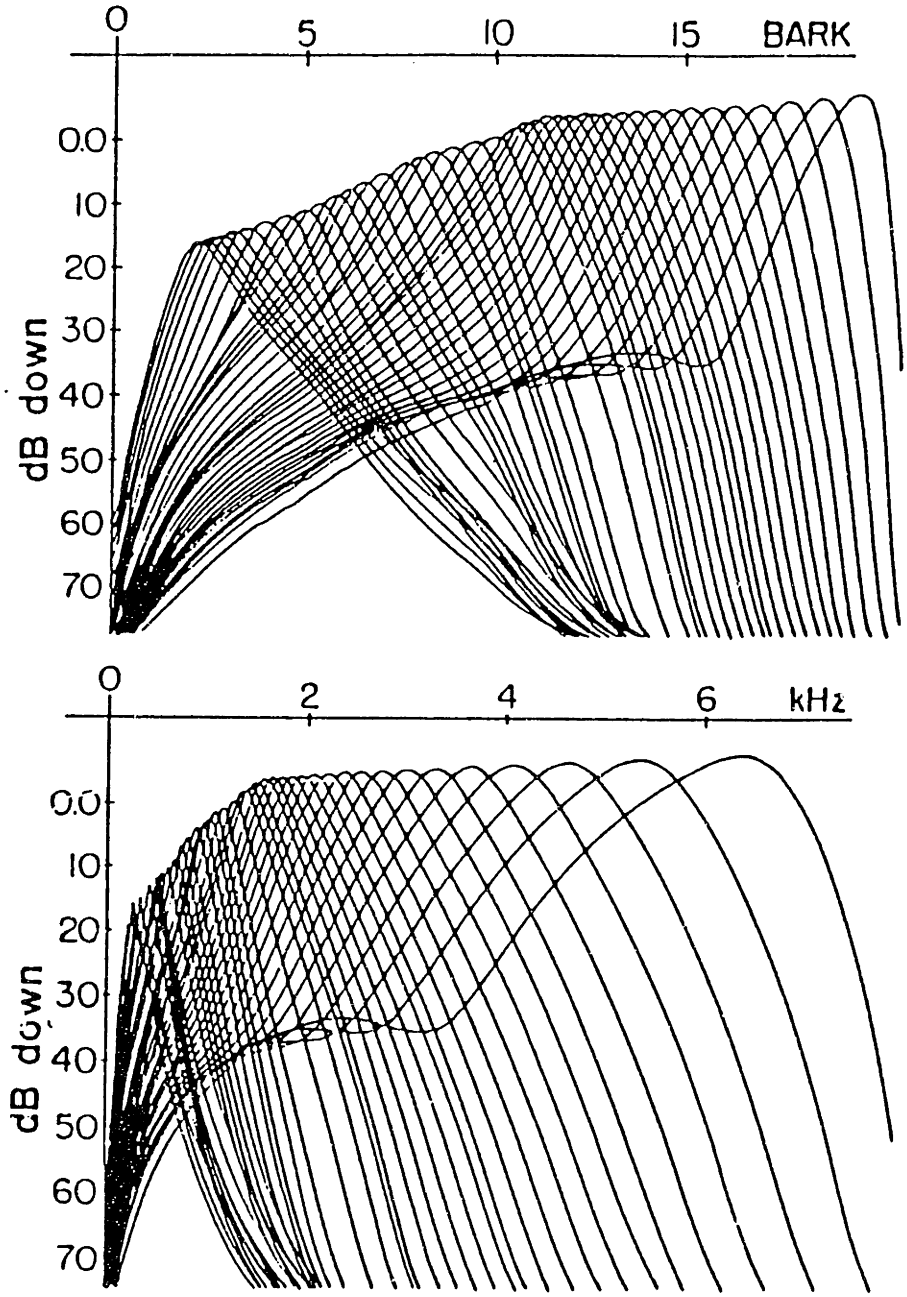

Figure 2.2: Frequency response characteristics of the critical band filter bank plotted along (a) a Bark scale and (b) a linear frequency scale (after Seneff)

Since the mean rate response (MR) and the synchrony response (SR) were intended to encode complementary acoustic information in the acoustic signal, a representation combining the two is also included in our experiments. This is done by appending the first 20 principal components [4] of the MR and SR to form another 40-dimensional vector (SAM-PC).

2.3.2 The Mel-frequency Representations

To obtain the mel-frequency spectral and cepstral coefficients (MFSC and MFCC, respectively), the signal is pre-emphasized via first differencing and windowed by a 25.6 ms Hamming window. A 256-point discrete Fourier Transform (DFT) is then computed from the windowed waveform. Following Mermelstein et al [16], these Fourier transform coefficients are later squared, and the resultant magnitude squared spectrum is passed through the mel-frequency triangular filter-banks described below. The log energy output (in decibels) of each filter, $X_k, k = 1, 2, .., 40$, collectively form the 40-dimensional MFSC vector. Carrying out a cosine transform on the MFSC according to the following equation yields the MFCC's, $Y_i, i = 1, 2, .., 40$.

$$Y_i = \sum_{k=1}^{40} X_k \cos[i(k - \frac{1}{2})\frac{\pi}{40}]$$

Some details about the cosine transform are provided in Appendix A. The lowest cepstrum coefficient, C_0, is excluded to reduce sensitivity to overall loudness.

In order to achieve as fair a comparison as possible, the mel-frequency triangular filter banks are designed to resemble the critical band filter bank of SAM (see Figure 2.3). The filter bank consists of 40 overlapping triangular filters spanning the frequency region from 130 to 6400 Hz. Thirteen triangles are evenly spread on a linear frequency scale from 130 Hz to 1 kHz, and the remaining 27 triangles are evenly distributed on a logarithmic frequency scale from 1 kHz to 6.4 kHz, where each subsequent filter is centered at 1.07 times the previous filter's center frequency. Since the bandwidths of the triangular

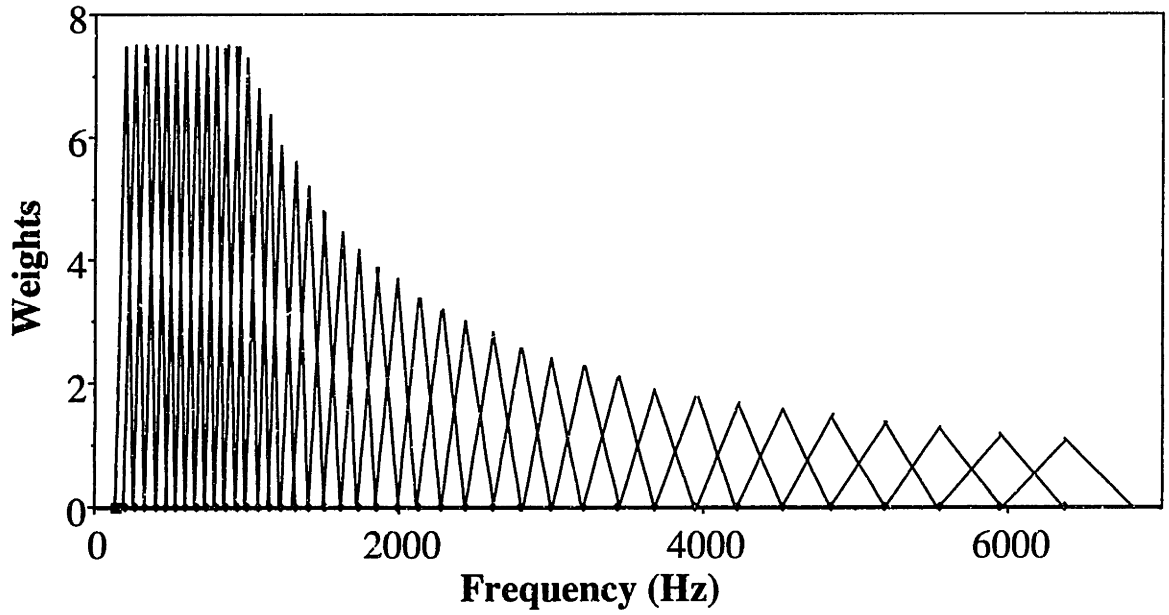

Figure 2.3: Design of the mel-frequency triangular filters

filters increase with the center frequencies, the area of each filter is normalized to unit magnitude in order to avoid amplification of the higher frequency coefficients through bandpass summation [26].

2.3.3 The Discrete Fourier Transform

To obtain the Fourier Transform representation, a DFT is computed in the same manner as described previously. Cepstral smoothing is performed to obtain a 256-point DFT, which is then down-sampled to 40 points. This processing sequence serves to filter out some non-essential pitch information.

2.3.4 Noise

One of the experiments which will be described below investigates the relative immunity of each representation to additive white noise. The noisy test tokens are constructed by adding white noise to the signal to achieve a peak signal-to-

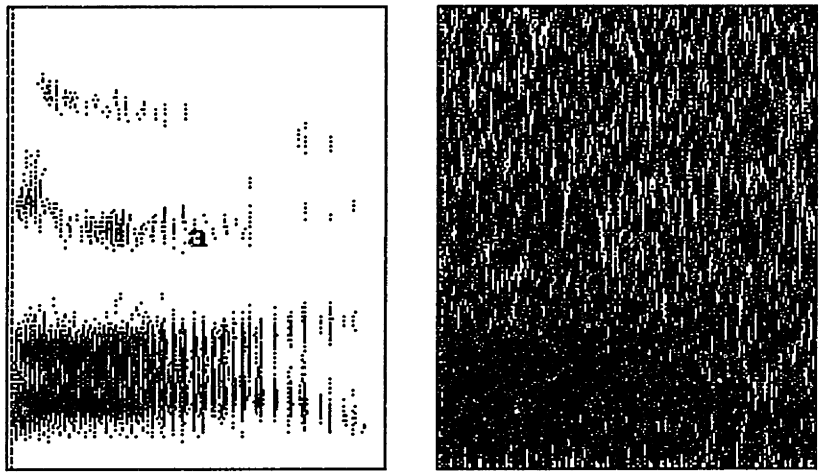

Figure 2.4: Wideband spectrograms showing clean and noisy speech for the vowel /a/

noise ratio (computed with the maximum energy in a frame of an utterance) of 20dB, which corresponds to a signal-to-noise ratio (computed with average energies) of slightly below 10dB. Figure 2.4 shows wideband spectrograms of one of the test tokens before and after noise corruption, and Figure 2.5 shows the corresponding spectra at the midpoint of the vowel token.

2.4 Task and Corpus

Comparisons of the various signal representations are based on the task of classifying 16 American English vowels using tokens excised from the acoustic-phonetically compact portion of the TIMIT database [19]. It is a classification task in that the boundaries of the vowel tokens are provided by the time-aligned phonetic transcription, and the classifier is only asked to determine the most likely label. The 16 vowels include 13 monophthongs /i, ɪ, e, ɛ, æ, a, o, ʌ, ɔ, u, ʊ, ü, ɝ/ and 3 diphthongs /ɑʸ, ɔʸ, ɑʷ/. No restrictions were imposed on the phonetic contexts in which they may appear. The training data consist of over 20,000 tokens, excised from 2,500 continuous sentences spoken by 500 speakers. The testing data consist of nearly 2,000 tokens, excised from 250

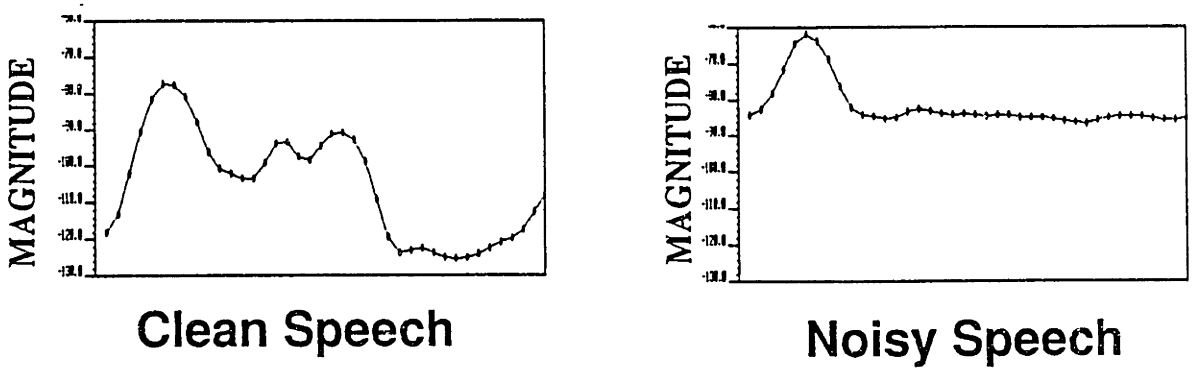

Figure 2.5: Smoothed DFT spectra of both clean and noisy speech for the midpoint of the vowel /a/

Training Speakers (M/F)	Testing Speakers (M/F)	Training Tokens	Testing Tokens
500 (357/143)	50 (33/17)	20,519	1,879

Table 2.1: Corpus used for the experiments

continuous sentences spoken by 50 new speakers. The size and contents of the corpus are summarized in Table 2.1.

2.5 The Artificial Neural Network Classifier

The classifier used for our experiment is an artificial neural network based on multi-layer perceptrons (MLP) [29]. The particular MLP architecture for phonetic recognition has previously been described in great detail by Leung [21]. The MLP is found to have several characteristics which are particularly advantageous for phonetic classification tasks, and in some cases, especially suited to our investigation. First of all, unlike a Gaussian classifier, for example, it does not make assumptions about the underlying probability distribution of the input data. Therefore, the classification performance is not penalized by

any invalid assumptions of the true underlying distributions.

Second, the MLP utilizes the training of connection weights to form decision regions, instead of using specific distance metrics (such as the Euclidean or Itakura [17]) to measure similarity. For traditional classifiers which do not assume probability distributions, the choice of a distance metric may be critical for robustness and performance [18]. Also, the distance metric may pose constraints on the input representation of a classifier. For example, discrimination by the Euclidean distance relies on differences in energy in the speech signal, and may be less suited for representations such as the synchronous response of SAM which has its energy information normalized. Since the experiments reported here involves several different acoustic representations, the MLP is particularly suitable for our purposes.

Third, the MLP accepts both continuous inputs such as acoustic attributes and/or binary inputs like linguistic features. This property, together with the two mentioned above, allows us to integrate heterogeneous sources of information as an input representation, as in the SAM-PC representation in our experiments.

Fourth, classification by the MLP is done through maximizing the differences between different classes by focusing on errors made at the decision surfaces, i.e. minimizing an error criterion. This is in contrast to the approaches which model individual classes independently of others, and may potentially be more effective in improving classification performance.

Fifth, the MLP is capable of forming disjoint decision regions in the multi-dimensional input space for the same class without supervision. This may be especially suitable for modelling the various allophones of a phoneme.

Finally, the MLP can be used as a hetero-associator to associate pairs of patterns. It is capable of mapping the complex speech signal to different levels of phonological and/or phonetic representations. Therefore, it can allow us to perform phonetic classification experiments as well as feature mapping experiments, as described in the next chapter.

2.5.1 Network Structure

The network used in this thesis has one hidden layer, and is illustrated in Figure 2.6. The number of output units N_0 depends on the number of classes to be recognized. In this case, there are 16 output units in our network, corresponding to the 16 vowels. The size of the network is determined by the number of units in the hidden layer, N_H. The number of input units N_I depends on the amount of input information available. In our experiments, the average spectra corresponding to the initial, middle and final third of the vowel token are appended together to form a 120-dimensional feature vector and used as input. This is done to implicitly capture the context dependency of vowel articulation. The inputs are normalized in amplitude and the connection weights are center initialized for better learning capabilities [20]. During supervised training, the inputs are fed forward through the network and the connection weights are updated for each training token to minimize a weighted mean squared error criterion. Details of the training and testing algorithm as well as previously improved parameters such as the number of hidden units that are used here have all been described in [21].

2.6 Results

For each acoustic representation, four separate experiments were conducted using 2,000, 4,000, 8,000, and finally 20,000 training tokens. In general, classification performance improves as more training tokens are utilized. This is illustrated in Figure 2.7, in which we display test set accuracies for the six different acoustic representations, using 2,000 and 20,000 training tokens. Each data point of test set accuracy is the average of 6 iterations, and the fluctuations between successive iterations are around 1%. The rest of the statistics are included in the Appendix B. For a fully trained network, the classification accuracies for different acoustic representations differ by about 5%, with the auditory-based representations consistently yielding better results than oth-

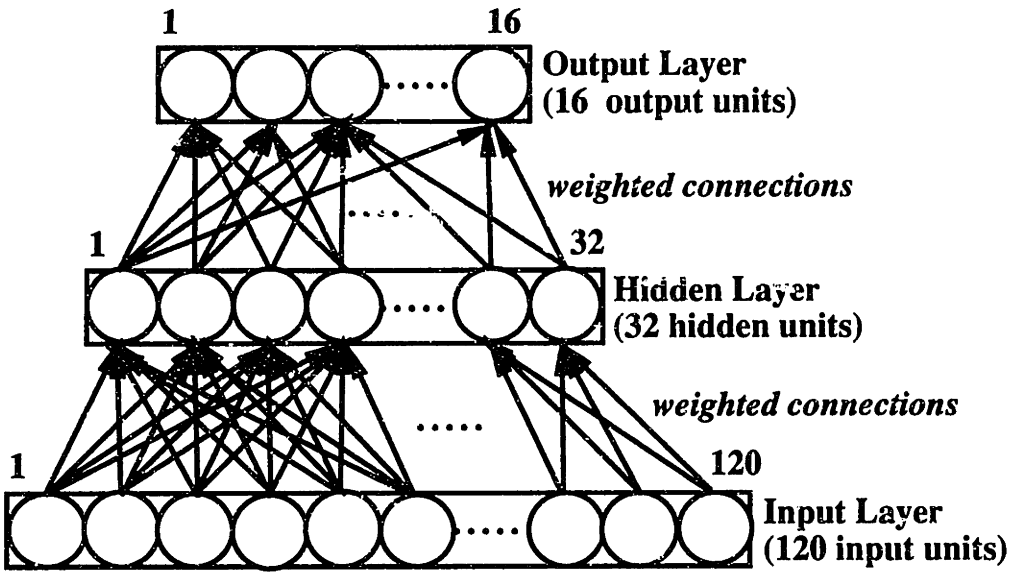

Figure 2.6: Structure of the Multi-Layer Perceptron Classifier

ers. According to a significance level of 0.01 using McNemar's test [9], the differences in performance of SAM-PC over each of the remaining representations are statistically significant, but this does not apply to the differences in performance between the remaining pairs of representations, as illustrated in Table 2.2.

In order to get some ideas about the robustness of the various representations, we also determined for each experiment the classification performance

	SAM PC	Mean Rate	Synchrony	MFSC	MFCC	DFT
SAM PC		SAM PC	SAM PC	SAM PC	SAM PC	SAM PC
Mean Rate			same	same	same	same
Synchrony				same	same	same
MFSC					same	same
MFCC						same
DFT						

Table 2.2: Results of McNemar's test on the performance of different acoustic representations (significance level = 0.01).

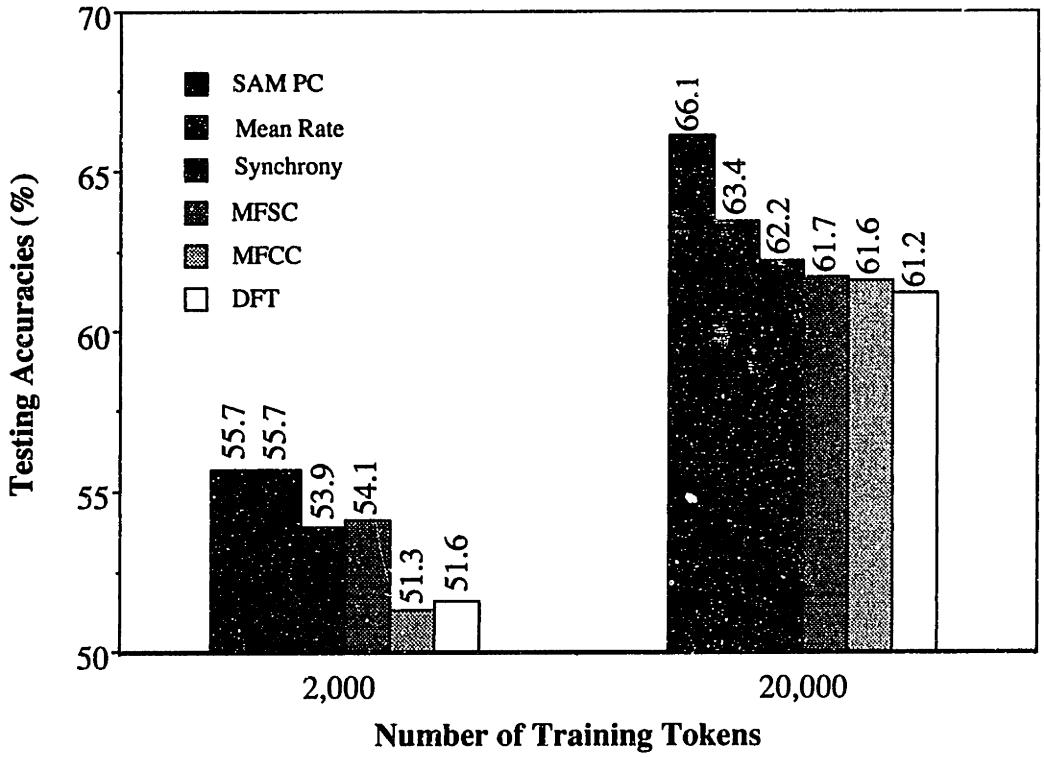

Figure 2.7: Performance of the six signal representations for 2,000 and 20,000 training tokens

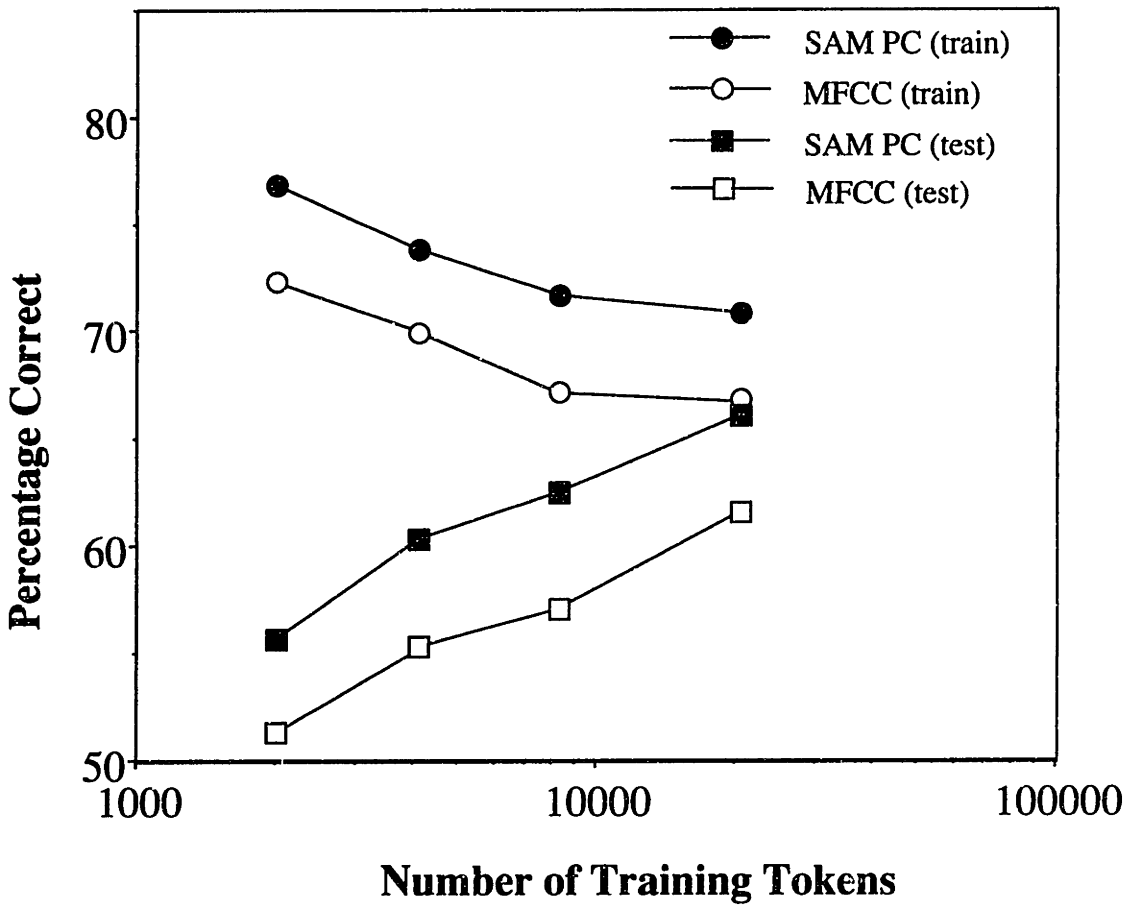

Figure 2.8: Effect of increasing training data on testing accuracies

on training data. Figure 2.8 shows accuracies on training and testing data as a function of the amount of training tokens for the combined auditory representation and the popular mel-frequency cepstral coefficients. As the size of the training set increases, so does the classification accuracy on testing data. This is accompanied by a corresponding decrease in performance on training data. At 20,000 training tokens, the difference between training and testing set performance is about 5% for both representations.

To investigate the relative immunity of the various acoustic representations to noise degradation, we determine the classification accuracy of the

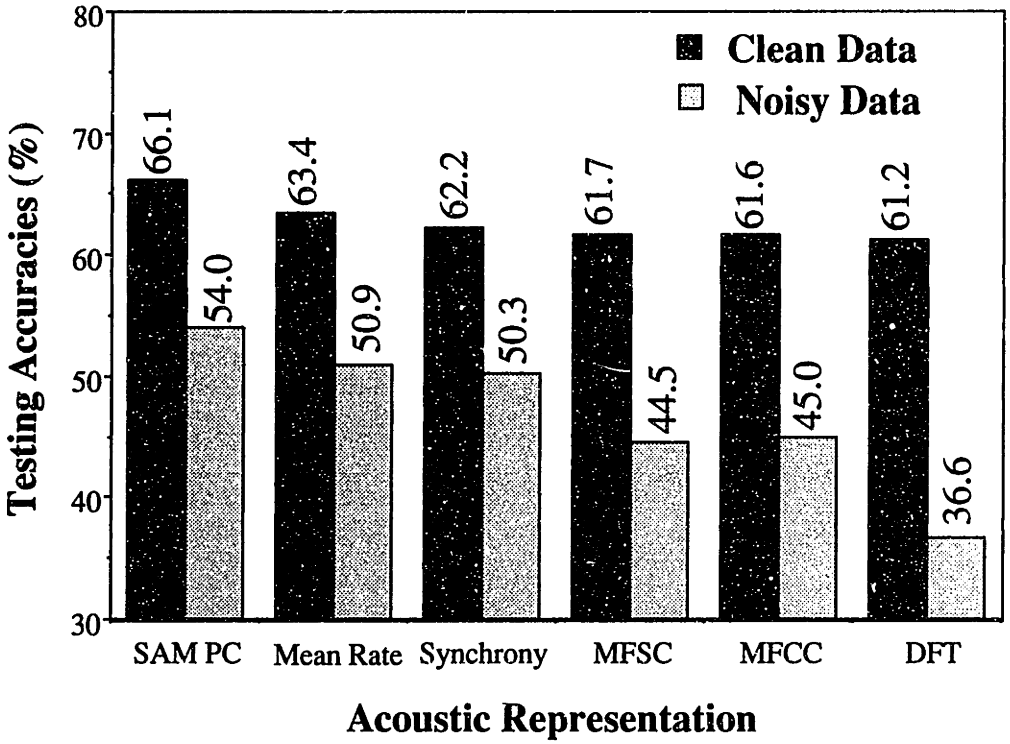

Figure 2.9: Performance of the different representations on noisy speech

noise-corrupted test set on the networks after they have been fully trained on clean tokens. The results with noisy test speech are shown in Figure 2.9, together with the corresponding results on the clean test set. The decrease in classification accuracy ranges from about 12% (for the combined auditory model) to almost 25% (for the DFT).

2.7 Discussion

Our results indicate that, on a fully trained network, acoustic representations based on auditory modelling consistently outperform other representations. The best among the three auditory-based representations, SAM PC, achieved a top-choice accuracy of 66%, which is comparable to those reported in the literature. For example, Leung [21] reported a classification accuracy of 64%,

with the same network and the same data set, when synchrony and mean-rate responses were used without principal component analysis.

When the two outputs of SAM are used separately, the performance typically drops by 3-4%, with the mean-rate response performing better than the synchrony response. This result is somewhat surprising, since the generalized synchrony detector (GSD) in SAM has the property of enhancing spectral peaks, whose locations are important for correct vowel identification. Apparently the mean-rate response also preserves the necessary acoustic information for vowel identification. It is also possible that the GSD algorithm over-sharpens the peaks in some cases, thus making the network unduely sensitive to amplitude variations at formant locations. Furthermore, the synchrony response lacks energy information, and cannot therefore distinguish as well between inherently louder vowels such as /**a**/ and other softer vowels such as /**u**/.

The MFSC and MFCC representations performed similarly on the fully trained network, worse than the auditory-based representations and slightly better than the DFT. At first glance, it may appear that the discrepancies are small, since the error rate is only increased slightly (from 33% to 38%). However, previous research on *human* and machine identification of vowels, *independent* of context, have shown that the best performance attained is around 65% [27]. Looking in this light, the difference in performance becomes much more significant.

One legitimate concern may be that principal component analysis has been applied to SAM PC, but not to MFCC. However, the cosine transform used in obtaining the MFCC perform a similar function as principal component analysis. To ensure that a fair comparison has been made, we have also conducted experiments in which principal component analysis is used on the MFCC. Taking 40 principal components as input yielded an average performance of 61.2%, which demonstrates that principal component analysis does not further improve the performance of the MFCC.

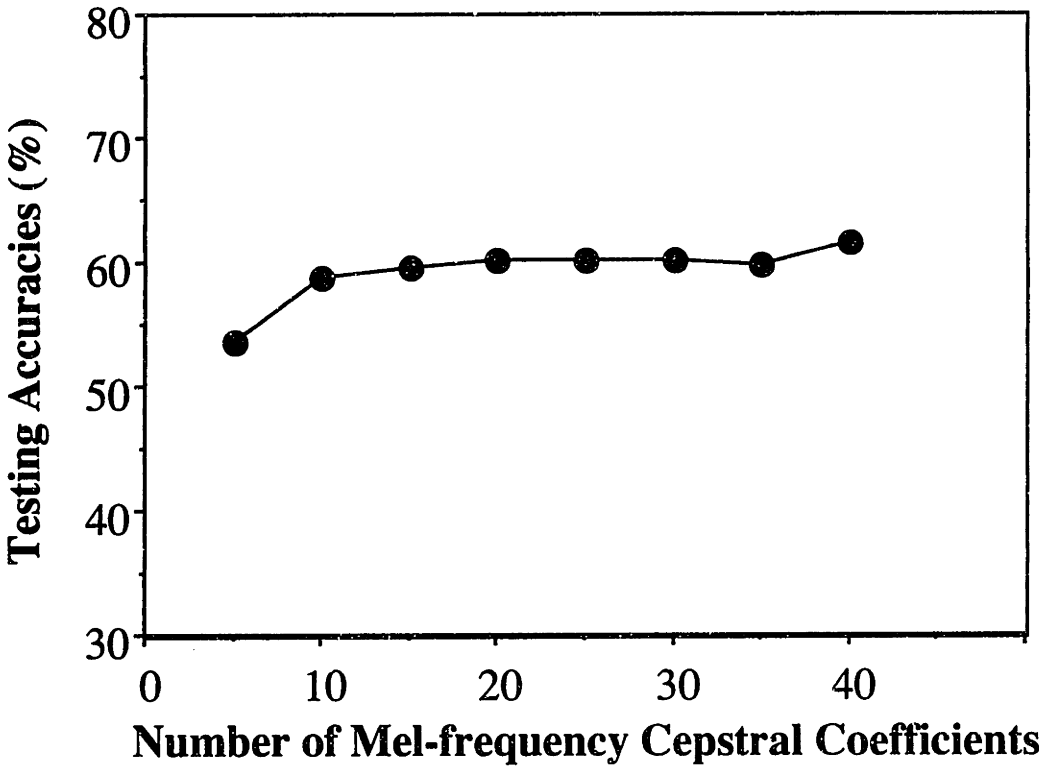

Figure 2.10: Effect of Varying the Number of MFCC on Vowel Classification Performance

Another concern may be that too many MFCC have been used. The higher order coefficients carry higher frequency spectral information, which is essential for vowel classification. So, using a large number of MFCC may to a certain extent cause classification performance to degrade. To resolve this issue, experiments have been performed where the number of MFCC used for the same vowel classification task is gradually increased from 5 to 40. Results are graphed in Figure 2.10, which shows classification performance does not decrease as more MFCC are used. Therefore, we may conclude that auditory-based signal representations are preferred, at least within the bounds of our experimental conditions.

As illustrated in Figs 2.7 and 2.8, the relative performances of the six rep-

resentations remained fairly stable as more training data were used. Overall, classification accuracy improved by an average of 9% as the training data increased ten-fold. The accuracies on the training set, on the other hand, decrease as expected with more training, suggesting that the network began to abstract relevant acoustic cues for phonetic distinction, rather than memorizing individual differences among tokens. The accuracies converge to less than 2% for DFT and over 5% for SR. If we regard the convergence between accuracies on the training and test sets as an indication of the increasing robustness of the network, then we can see from Figure 2.8 that for different acoustic representations, the robustness is increasing at approximately the same rate. With additional training data, we would expect that the test set accuracy can continue to improve. However, it is not very likely that relative performances will change.

In the presence of noise, classification performance degraded for all the representations. While the relative performances follows the trend of clean speech, the differences between different representations varied substantially. The degradation of the SAM representations was least severe - about 12%, whereas the mel-representations showed a drop of 17%. The DFT is most affected by noise, and its performance degraded by over 24%. Figure 2.11 shows the clean and noisy versions of the same vowel token shown in Figure 2.4. The respective spectra at the mid-point of the vowel token are shown in Figure 2.12. The fact that the SAM representations are more immune to noise can be gleaned from comparing Figure 2.11 with Figure 2.4, and comparing Figure 2.12 with Figure 2.5. Most of the formant information in the noisy signal is preserved in the synchrony response, but such information is difficult to detect in the DFT.

We believe that training with clean speech and testing with noisy speech is a fair experimental paradigm since the noise level of test speech is often unknown in practice, but the environment for recording training speech can always be controlled.

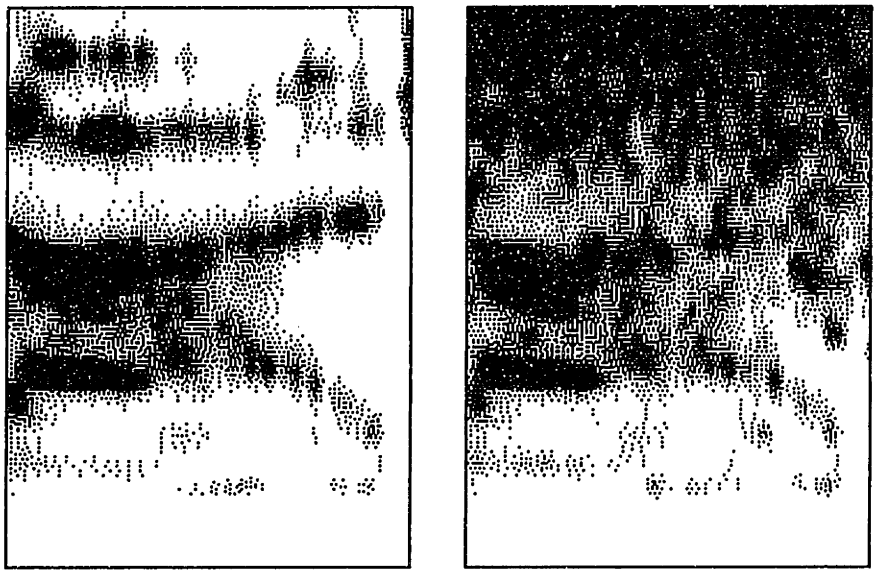

Figure 2.11: Synchrony spectrograms showing clean and noisy speech for the vowel /a/

2.8 Chapter Summary

In this chapter, we reported the results of a set of vowel classification experiments that compare the relative merits of six acoustic representations. We found that, for clean testing tokens, the auditory based representations hold a small but consistent advantage over the other representations. This advantage is magnified greatly when the testing tokens are corrupted by noise. In the following chapter, we will be pursuing other issues related to the acoustic to lexical transformation. Specifically, we would like to determine whether one should use the signal representation directly, or attempt to extract acoustic attributes that may better signify phonetic contrasts. We will also explore the possibility of introducing distinctive features as an intermediate lexical representation. The auditory models will be used for all of these experiments.

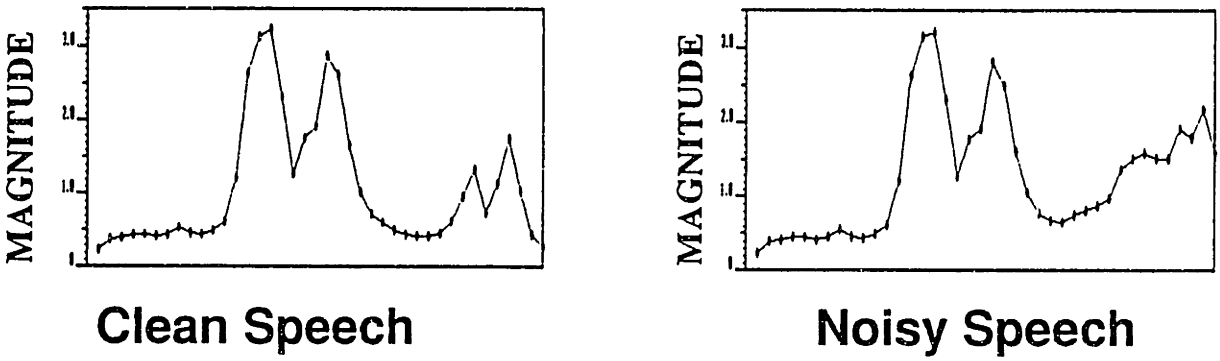

Figure 2.12: Synchrony spectra showing clean and noisy speech for the mid-point of the vowel /a/

Chapter 3

Attribute Extraction and Distinctive Features

In this chapter, we continue our study by focusing on the two remaining questions: "Should we use spectral representation directly for phoneme/feature classification, or should we extract and use acoustic attributes instead?" Furthermore, does the introduction of an intermediate feature-based representation between the signal and lexicon offer performance advantages?" We have chosen to answer these questions by performing a set of phoneme classification experiments in which conditional variables are systematically varied. The usefulness of one condition over another is inferred from the performance of the classifier.

3.1 Experimental Paradigm

We have mentioned in Chapter 1 (Figure 1.2) that it is uncertain how we should utilize distinctive features in our speech decoding strategy. We can extract acoustic attributes based on distinctive features, and use the attributes to replace the direct use of the spectral representation. We can also implement an intermediate phonological representation between the signal and the lexicon based on distinctive features. It is not clear which method we should use, or whether we should use both. Algorithms need to be designed for extracting acoustic attributes, for mapping the acoustics to the intermediate phonological

representation, as well as bridging the gap between the intermediate representation and the lexicon. In this chapter, we describe an experimental paradigm designed to compare the various possible pathways of speech decoding. Three experimental parameters were systematically varied, resulting in six different conditions, as depicted in Figure 3.1. These three parameters specify whether the acoustic attributes are extracted, whether an intermediate distinctive feature representation is used, and how the feature values are combined for vowel classification.

In some conditions (cf. conditions A, E, and F), the spectral vectors were used directly, whereas in others (cf. conditions B, C, and D), each vowel token was represented by a set of automatically-extracted acoustic attributes. In still other conditions (cf. conditions C, D, E, and F), an intermediate representation based on distinctive features was introduced. The feature values were either used directly for vowel identification through one bit quantization (i.e. transforming them into a binary representation) followed by table look-up (cf. conditions C and E), or were fed to another MLP for further classification (cf. conditions D and F). Our experiments were again conducted using an MLP classifier for speaker independent vowel classification. Taken as a whole, these experiments will enable us to answer the questions that we posed earlier. For example, we can assess the usefulness of extracting acoustic attributes by comparing the classification performance of conditions A versus B, D versus F or C versus E. Each of these three pairs show the contrast between using the spectral representation directly and extracting and using acoustic attributes. To assess the usefulness of incorporating an intermediate feature-based representation, we can compare conditions B versus C, or B versus D. These results should be corroborated by comparing conditions A versus E and A versus F respectively. As for assessing the effectiveness of feature classification, we can compare conditions C versus D, and E versus F, and it is expected that the two comparisons should yield similar observations.

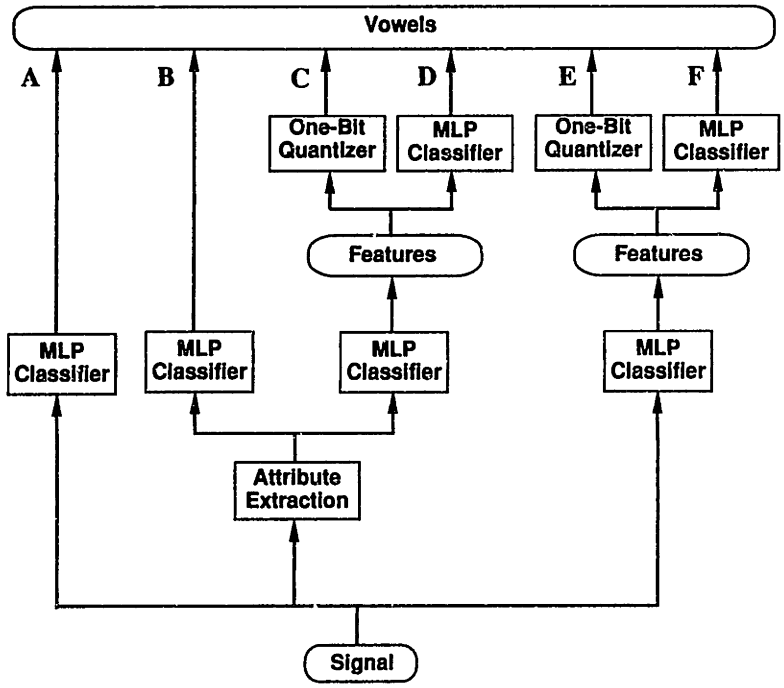

Figure 3.1: Experimental paradigm comparing direct phonetic classification with attribute extraction, and the use of linguistic features.

Training Speakers (M/F)	Testing Speakers (M/F)	Training Tokens	Testing Tokens
500 (357/143)	50 (33/17)	18,558	1,672

Table 3.1: Corpus used for the experiments

3.2 Task and Corpus

The task chosen for our experiments is the classification of 13 monophthong vowels in American English – /i, ɪ, e, ɛ, æ, a, o, ʌ, ɔ, u, ʊ, ü and ɝ/. The diphthongs are excluded here because their dynamic nature may render distinctive feature specification ambiguous. Consequently, there are fewer training and testing tokens compared with our previous corpus (cf. Table 3.1).

Following the conventions set forth by others [37], we characterized the 13 vowels in terms of 6 distinctive features. The feature values for these vowels are summarized in Table 3.2.

	i	ɪ	e	ɛ	æ	a	ɔ	o	ʌ	u	ɝ	ʊ	ü
HIGH	+	+	-	-	-	-	-	-	-	+	-	+	+
TENSE	+	-	+	-	-	-	-	+	-	+	-	-	+
LOW	-	-	-	-	+	+	+	-	-	-	-	-	-
BACK	-	-	-	-	-	+	+	+	+	+	+	+	-
ROUND	-	-	-	-	-	-	+	+	-	+	+	+	+
RETROFLEX	-	-	-	-	-	-	-	-	-	-	+	-	-

Table 3.2: The Set of Distinctive Features used to characterize 13 vowels

3.2.1 Spectral Representation

The spectral representation is obtained from Seneff's auditory model, since its representations have been found to be superior to others during our previous study [23]. While the combined mean rate and synchrony representation (SAM-PC) gave the best performance, it may not be an appropriate choice for

our present work, since the heterogeneous nature of the representation poses difficulties in acoustic attribute extraction. As a result, we have selected the next best representation - the mean rate response (MR). This representation consists of 40 spectral coefficients spaced half bark apart and computed every 5 ms. A 120-dimension feature vector is obtained by appending the three average vectors representing the input token.

3.2.2 Acoustic Attributes

Each vowel token is characterized either directly by a set of spectral coefficients, or indirectly by a set of automatically derived acoustic attributes. In the latter case, the attributes that we extract are intended to correspond to the acoustic correlates of distinctive features. However, we are confronted with several problems. First, we do not as yet possess a full understanding of these correlates for each feature. Even in cases where these correlates have been proposed, they are typically described in terms of parameters such as formant frequencies, which are obtained through heuristic methods and can lead to catastrophic measurement errors. Besides, we must somehow capture the variabilities of these features across speakers and phonetic environments. For these reasons, we have adopted a more statistical and data-driven approach. In this approach, a general property detector is proposed, and the specific numerical values of its free parameters are determined from training data using an optimizatic criterion proposed by Phillips [38]. In our case, the general property detectors chosen are the spectral center of gravity and its amplitude. This class of detectors may carry formant information, and can be easily computed from a given spectral representation. As discussed previously, the mean rate response is used.

The process of attribute extraction is as follows. First, speaker normalization is done by shifting the spectrum down linearly on the bark scale by the median pitch [32]. Then, for each distinctive feature, the training tokens are divided into two classes: [+feature] and [-feature]. The lower and upper

frequency edges (or "free parameters") of the spectral center of gravity are chosen so that the resultant measurement can maximize the Fisher's Discriminant Criterion (FDC) between the classes [+feature] and [-feature]. The FDC is defined as the ratio of the difference in class means and the total within-class scatter of the samples. It is given by the following formula: [5]

$$J(x,y) = \frac{|m_1(x,y) - m_2(x,y)|^2}{s_1(x,y)^2 + s_2(x,y)^2},$$

where x, y are the lower and upper frequency edges used to compute the spectral center of gravity, $m_1(x,y)$ and $m_2(x,y)$ are the means of centers of gravity for the classes [+feature] and [-feature] respectively, and $s_1(x,y)^2$ and $s_2(x,y)^2$ are the variances of centers of gravity for the classes [+feature] and [-feature] respectively.

For the features [BACK], [TENSE], [ROUND], and [RETROFLEX] only one attribute per feature is used. For [HIGH] and [LOW], we found it necessary to include two attributes per feature, using the two sets of optimized free parameters giving the highest and the second highest FDC. These 8 frequency values, together with their corresponding amplitudes, make up 16 attributes for each third of a vowel token. Therefore, performing acoustic attribute extraction has the effect of reducing the input dimensions from 120 to 48. The specific attributes used are included in Appendix C.

3.3 Classification Procedures

The classifier used for our experiments here is again the MLP with a single hidden layer with 32 hidden units. As can be seen from Figure 3.1, some of the MLP's classify the input directly into one of 13 vowels and therefore possess 13 output units. The others map the input into an intermediate representation of distinctive features. In this case, the output consists of six units, each corresponding to some probability measure of the accurate mapping of a distinctive feature.

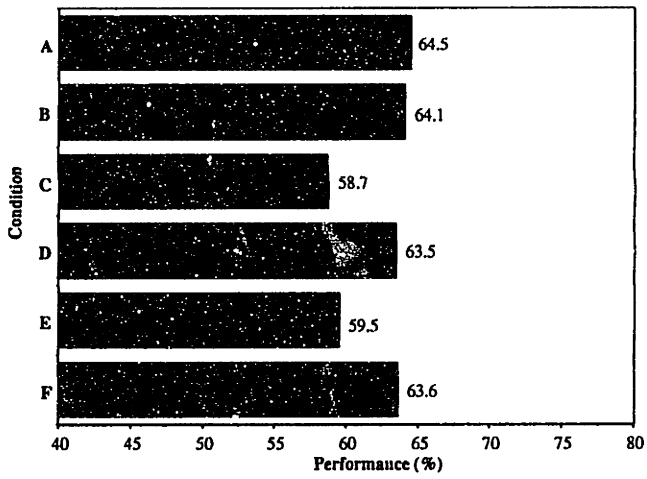

Figure 3.2: Performance of the six classification pathways in our experimental paradigm

The structures of all the above networks are similar to that used in the signal representation experiments. Each has a single hidden layer with 32 hidden units. Once again, input normalization and center initialization have been used [20].

3.4 Results

The results of our experiments are summarized in Figure 3.2, plotted as vowel classification accuracy for each of the conditions shown in Figure 3.1. The values in this figure represent the average of 6 iterations; performance variation among iterations of the same experiment amounts to about 1%.

Comparing the results for conditions A and B, we found no statistically significant difference in performance, according to McNemar's test, as we replace the spectral representation by the acoustic attributes (see Table 3.3). This result is further corroborated by the comparison between conditions C and E, and D and F.

Figure 3.2 shows a 4-5% deterioration in performance when one simply maps the feature values to a binary representation for table look-up (i.e., com-

paring conditions A to E and B to C). This deterioration is statistically significant (Table 3.3). We can also examine the accuracies of binary feature assignment for each feature, and the results are shown in Figure 3.3. The accuracy for individual features ranges from 87% for [ROUND] and [TENSE] to 98% for [RETROFLEX], and there is again little difference between the results using the mean rate response and using acoustic attributes. It is perhaps not surprising that table look-up using binary feature values results in lower performance, since it would require that *all* of the features be identified correctly.

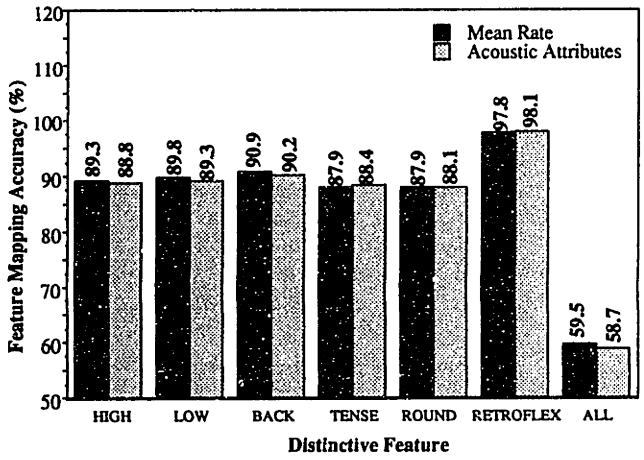

Figure 3.3: Distinctive Features Mapping Accuracies for the Mean Rate Response and Acoustic Attributes

However, when we use a second MLP to classify the features into vowels, a considerable improvement ($> 4\%$) is obtained to the extent that the resulting accuracy shows no significant difference from other conditions (cf. conditions A and F, and conditions B and D).

3.4.1 Significance Testing

Table 3.3 shows the result of McNemar's test comparing different conditions in the paradigm with the significance level of 0.001. The entries in the table may

either show the better condition, or indicate that the two conditions are the same. Essentially, there is no significant deterioration in performance as we replace the spectral representation with attributes, no significant deterioration in performance as we incorporate an intermediate feature-based representation, but significant deterioration if the feature values are quantized and then used for table-lookup.

	A	B	C	D	E	F
A		same	A	same	A	same
B			B	same	B	same
C				D	same	F
D					D	same
E						F
F						

Table 3.3: Results of McNemar's test comparing the six conditions in our paradigm (significance-level = 0.001)

3.5 Discussion

Our investigation on the use of acoustic attributes is partly motivated by the belief that these attributes can enhance phonetic contrasts by focusing upon relevant information in the signal, thereby leading to improved phonetic classification performance when only a finite amount of training data is available. The acoustic attributes that we have chosen are intuitively reasonable and easy to measure. But they are by no means optimum, since we did not set out to design the best set of attributes for enhancing vowel contrasts. Nevertheless, their use has led to performance comparable to the direct use of spectral information. With an improved understanding of the relationship between distinctive features and their acoustic correlates, and a little more care in the design and extraction of these attributes, it is conceivable that better classification accuracy can be obtained.

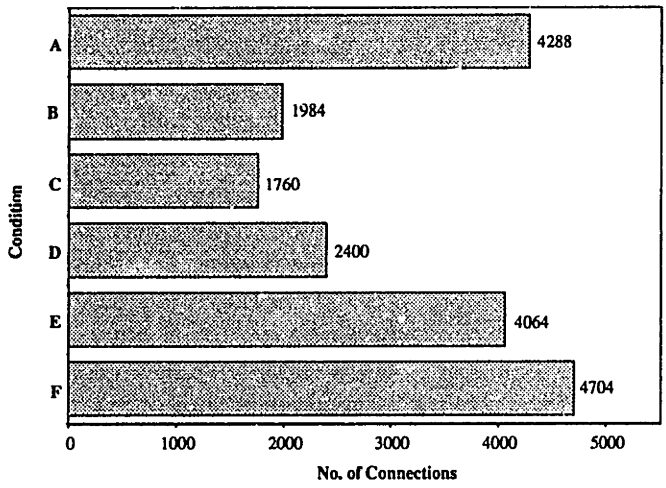

Figure 3.4: Network complexities of the various classification conditions in our experimental paradigm

Another advantage of using acoustic attributes is savings on run-time computations through reduction of input dimensions. Figure 3.4 compares the complexities, measured as the number of connections in the artificial neural network, for each condition in our experimental paradigm. With a small amount of preprocessing for computing the attributes, the use of acoustic attributes can save about half of the computations required by the direct use of spectral representation.

One potential source of discrepancy in our experiments has to do with pitch normalization. No pitch normalization was performed on the mean-rate response, whereas a pitch-normalized spectral center of gravity measure was used as acoustic attributes. Pitch normalization in attribute extraction was thought to be desirable since it can eliminate singularities that complicate the search for a maximum FDC value in the optimization process as illustrated in Figure 3.5, which plots the FDC score on the z-axis, and the lower and upper frequency edges x and y on the x- and y- axes respectively. The frequency edges yielding the highest FDC score are selected as the "optimized" free parameters, as illustrated in Figure 3.5. The global maximum is easy to

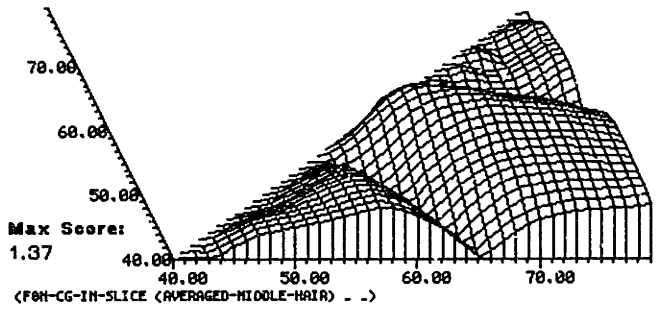

Figure 3.5: Choosing lower and upper frequency edges for the spectral center of gravity to represent the feature BACK

find in this case since the three-dimensional surface is smooth. However, if pitch normalization has not been included in our attribute extraction process, "spikes" may appear on the three-dimensional surface. These spikes have high FDC values regardless of the contour of the surface, i.e. they may be located at local minima. Therefore, we have chosen to include pitch normalization in our optimization process. We have conducted further experiments where pitch normalization is included in the conditions A, E and F, and the performance improvement obtained is below 1.5% in each case. According to McNemar's test with a significance level of 0.001, the difference in performance is not statistically significant. Therefore, any performance advantages that may be brought about by speaker normalization is not an issue.

To introduce a set of linguistically motivated distinctive features as an intermediate representation for phonetic classification, we first transform the acoustic representations into a set of features, and then map the features into vowel labels. While one may argue that such a two-step process is inherently sub-optimal, we nevertheless were able to obtain comparable performance, corroborating the findings of Leung [21]. Such an intermediate representation can offer us a great deal of flexibility in describing contextual variations. For example, all vowels sharing the feature [+ROUND] will affect the acoustic properties of neighboring consonants in predictable ways, which can be described

more parsimoniously. By describing context dependencies this way, we can also make use of training data more effectively by collapsing all available data along a given feature dimension.

Figure 3.3 shows that performance on some features is worse than others, presumably due to inadequacies in the attributes that we use. For example, performance on the feature [TENSE] should be improved by incorporating segment duration as an additional attribute. When a second classifier is used to map the feature values into vowel labels, a 4-5% accuracy increase is realized such that the performance is again comparable to cases without this intermediate feature representation. This result suggests that the acoustic-phonetic information is preserved in the *aggregate* of the features, and that the subsequent performance recovery may be a consequence of the redundant nature of distinctive features, as well as the ability of the second classifier to capture various contextual effects.

3.6 Error Analyses

In order to compare the different experimental conditions in our paradigm more thoroughly, the classification errors made in a typical iteration of each condition A, B, D and F are tabulated in confusion matrices shown in Tables D.1 to D.4 of Appendix D respectively. The rows correspond to the stimulus to the network - the first entry of each row holds the transcription label of the input token, and the last entry is the total number of test tokens carrying that transcription. The columns correspond to the response of the network - the first entry of each column represents the vowel label assigned to the input token as a result of classification, and the last entry is the total number of test tokens being assigned that label. Each of the remaining entries is a percentage of vowel tokens. For example, in Table D.1, the fifth row and the sixth column together show that out of the 158 /ɛ/ vowels in the testing data, 16.5% have been mislabelled as /ɪ/, and there were a total of 230 test vowels labelled by

the classifier as /ɪ/.

3.6.1 Mutual Information

To measure the performance of each condition, we compute the mutual information between the random variable X of the transcription labels, and the random variable Y of the vowel labels produced by the network [21,8]. The mutual information measures the average reduction of uncertainty in X after observing Y, and is given by the equation :

$$I(X;Y) = H(X) - H(X|Y) \tag{3.1}$$

where $I(X;Y)$ is the mutual information between random variables X and Y, $H(X)$ is the entropy of X which measures its average uncertainty,

$$H(X) = -\sum_{x} P_X(x) log P_X(x) \tag{3.2}$$

and $H(X|Y)$ is the conditional entropy which measures the uncertainty in X having observed Y, given by

$$H(X|Y) = -\sum_{xy} P_{XY}(x,y) log P_{X|Y}(x|y) \tag{3.3}$$

In the above equations, $P_X(x)$ is the probability distribution of X, $P_{X|Y}(x|y)$ is the conditional probability distribution of X given Y and $P_{XY}(xy)$ is the joint probability of X and Y.

The mutual information is computed using the statistics from each confusion matrix, and the result is tabulated in Figure 3.6.

The mutual information for conditions C and E are not computed because some tokens have ambiguous feature assignments which do not match any feature set of the 13 vowels in our vocabulary. An example of an ambiguous feature vector is ([-HIGH], [-LOW], [-BACK], [-TENSE], [-ROUND] and [+RETROFLEX]. This feature vector originates from a test vowel /ɝ/, but with

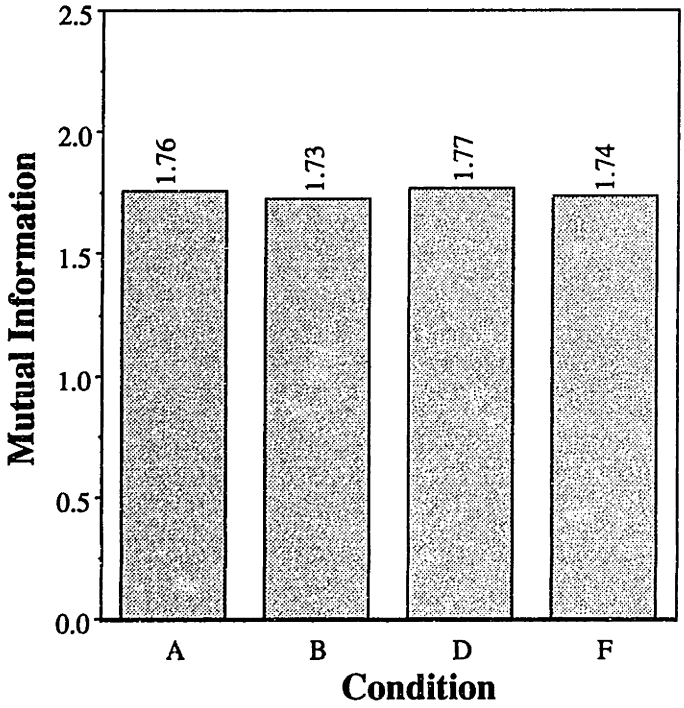

Figure 3.6: Mutual information computed from the confusion matrices of conditions A, B, D and F in the experimental paradigm

the two features [BACK] and [ROUND] mapped incorrectly. Across the remaining conditions, we obtain comparable mutual information values from the respective confusion matrices, which show that there is no loss of information caused by extracting acoustic attributes or implementing an intermediate feature-based representation.

3.6.2 Utility of Feature Classification

In this subsection, we will address the usefulness of incorporating a second MLP in the classification pathway, (c.f. conditions C and D, and conditions E and F). The first MLP in conditions C and E delivered a set of ambiguous features for over 5% of the test tokens, and therefore no vowel label could be assigned to the tokens by table lookup. Mistakes made in a typical iteration of the table-lookup procedure may be found from the confusion matrices in Tables D.5 and D.6 included in appendix D. For example, in both conditions C and E, a very prominent ambiguous feature vector is (001010) corresponding to the features (-HIGH, -LOW, +BACK, -TENSE, +ROUND, -RETROFLEX). This error occurs for a variety of input tokens, and most frequently for the phonemes /ɔ/ and /o/, which should have correct feature values of (011010) and (001110) respectively. Another example is the ambiguous feature vector (100010), which tends to occur to the phonemes /ɪ/ and /ü/ which should have correct feature values of (100000) and (100110) respectively. One of the causes of failure in the table-lookup procedure lies in the fact that it puts equal weighing for *all* the features characterizing a phoneme, whereas in actuality, a phoneme can be identified by accurate classification of *some* of the features. For instance, the vowel /u/ is often fronted when surrounded in alveolar context to form the vowel /ü/. Consequently, the feature [+BACK] is relatively unimportant in the identification of the vowel. The more crucial features are perhaps [+HIGH], [+TENSE] and [+ROUND]. In other words, in order to correctly classify the vowel /u/ from a set of feature outputs, we should weigh [+HIGH], [+TENSE] and [+ROUND] much heavier than [+BACK]. In addition, this set of weights

should only apply to /u/, and every other phoneme should have its own specific set of weights.

The second MLP classifier is better able to handle this problem. The connection weights have been trained so that when the feature set of a test token is fed forward in the network, some features are weighted more heavily than others. Therefore, conditions D and F are able to assign a vowel label to all the ambiguous feature sets which occur in conditions C and E. Moreover, the second classifier is also able to correct some of the the classification errors, although at other times, it may alter an originally correct decision. In the iteration of condition D, out of the 970 feature errors made by the first classifier, the second classifier corrected 189 but confused 160 originally correct features, resulting in 941 feature errors after the second classifier. In condition F, out of the 956 feature errors produced by the first classifier, the second classifier corrected 148 but confused 154 originally correct features, resulting in 962 feature errors after the second classifier. Despite an increase of 6 feature errors in the latter case, the second MLP classifier was able to recover the performance from 59.9% of the table-lookup procedure to 63.1%, which indirectly shows that some features are more important than others in the recognition of different phonemes and performance would not be affected as much if the feature mapping mistakes were made on the less crucial features or if the errors are correlated. For each vowel, we can compare its proper feature assignment from Table 3.2 with the quantized feature mapping output from the network. The number of features different between the proper feature set and the mapped feature set ranges from 0 (all features mapped correctly) to 6 (all features mapped incorrectly). The cumulative percentage of tokens is plotted against the number of binary features different, as shown in Figures 3.7 and 3.8. We can see from these plots that over 95% of the confusions occur between vowels that have two or fewer features different. Furthermore, comparing conditions C with D and E with F, we can see that there is an increase in the number of tokens with all features mapped correctly, and a slight decrease in the number

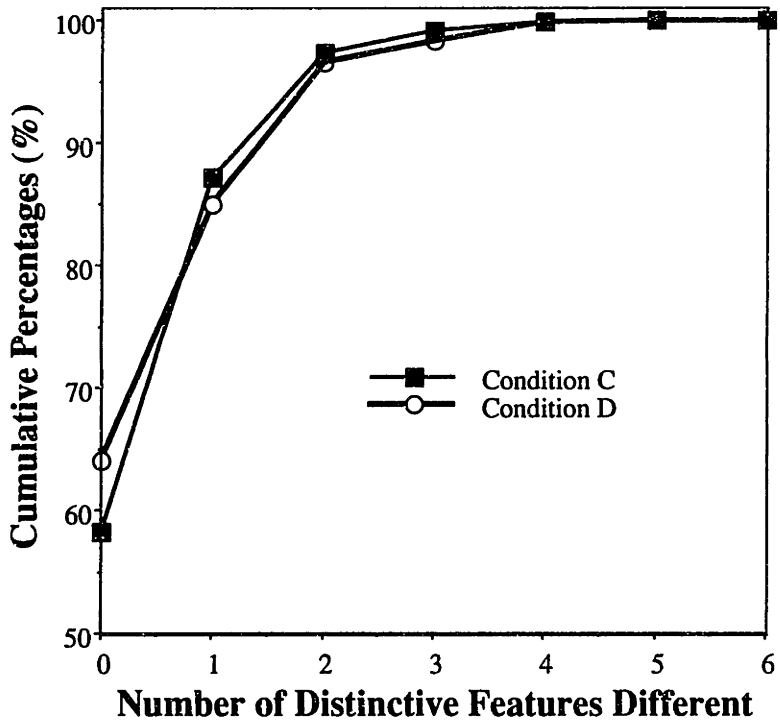

Figure 3.7: Performance of conditions C and D in terms of the number of features different between network outputs and transcription labels

of tokens with one or two binary features different, suggesting that the second classifier mostly corrects near-misses. For example, in conditions C and E where the table lookup method is used, some of the vowel tokens of /ü/ adjacent to the semivowel /l/ are often classified as /u/, while others adjacent to the semivowel /y/ are often classified by as /i/. Other examples include misclassifying /e/ or /ɪ/, as /ɛ/ or /i/, and misclassifying a nasalized /æ/ as /ɛ/. In these cases, the mistakes made are quite often corrected by the second MLP.

3.7 Chapter Summary

In this chapter, we have described a methodology to extract a reasonable set of acoustic attributes which attempts to capture the relevant aspects of the acous-

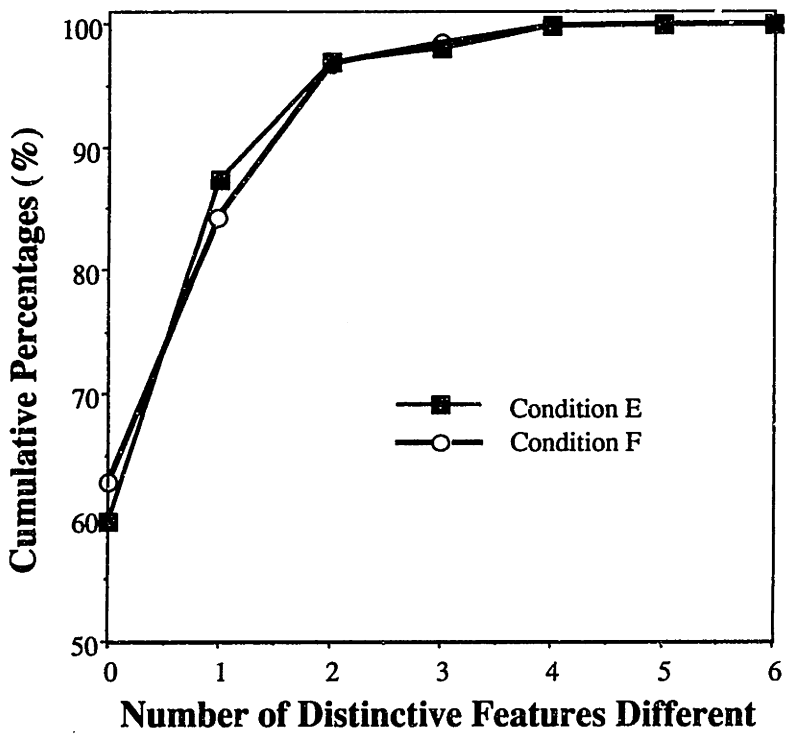

Figure 3.8: Performance of conditions E and F in terms of the number of features different between network outputs and transcription labels

tic signal for vowel classification. We have found that the use of such acoustic attributes can significantly reduce run-time computation for feature mapping and vowel classification with little cost to accuracy. Furthermore, the introduction of an intermediate representation based on distinctive features can potentially provide us with a flexible framework to describe contextual variations and make more effective use of training data, at no cost to classification performance.

Chapter 4

Signal Representation for Acoustic Segmentation

This chapter is a brief extension of the work reported in Chapter 2, where we make further comparisons of several acoustic representations based on segmentation. Our attention is focused upon delineating the acoustic signal into segments, where each segment correspond to an individual acoustic event. Such an event, or group of events, can eventually be mapped into phonemes. The vowel classification experiments reported in Chapter 2 use vowel tokens which have been excised from the original speech signal using a time-aligned phonetic transcription. Since phonetic recognition involves not only phonetic classification, but segmentation as well, we should also investigate the appropriateness of signal representations for this second task.

For this part of our investigation, we use an automatic procedure for acoustic segmentation previously developed by Glass [10], where acoustic events are embedded in a multi-level structure called a *dendrogram*. This method of segmentation has an advantage over others based on single-level descriptions, since it is capable of distinguishing fine to coarse acoustic changes in an utterance. Furthermore, dendrogram segmentation uses relative measures in the acoustic signal, which makes it more robust and largely independent of effects such as speaker characteristics and background noise. Previously, dendrogram segmentation has been used in conjunction with auditory models, where the

onsets and offsets of sounds tend to be sharpened [11,33]. In the following sections we will report experiments that have been conducted to compare different acoustic representations for dendrogram segmentation.

4.1 Signal Representations

The three spectral representations compared here include one from the Seneff's auditory model (the mean rate response), one mel-frequency representation (MFSC), and the smoothed DFT. Processing sequences of these three representations have previously been described in detail in Chapter 2. Segmentation for each representation was done using an array of feature vectors as input.

4.2 Acoustic Segmentation Algorithm

The algorithm used to establish acoustic segments is developed by Glass [11]. It aims to divide the acoustic signal into segments which are acoustically homogeneous. The procedure starts by measuring the similarity between each frame and its neighboring frames (10 ms away), using a distance metric. An association is then established between a given frame and its more similar neighbor, from left to right along the time axis. When the association switches from past to future, an acoustic boundary is marked.

The above procedure will divide a given utterance into many small segments. Such segments are used as "seed regions" and the average spectrum for each region is computed. Two regions are merged to form a new single region if they are more similar to each other than to the other neighboring region. This is done repetitively, with increasing distances between adjacent regions, until the entire utterance is described by a single region. The complete process for an utterance can be displayed in a *dendrogram* by plotting the distance between merged regions versus time, as illustrated in Figure 4.1. Boundaries closer to the bottom of the dendrogram describe finer acoustic transitions,

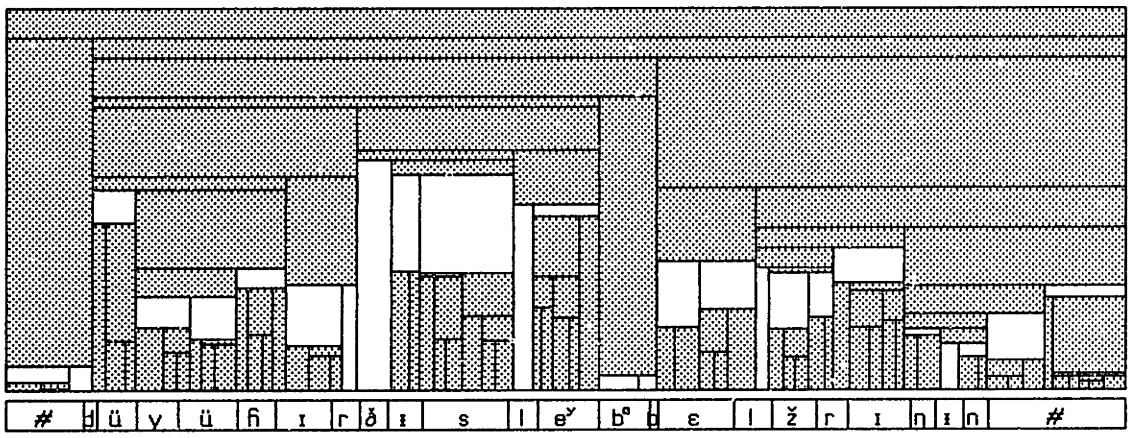

Figure 4.1: A dendrogram computed with a Euclidean distance metric.

whereas those nearer the top describe more abrupt acoustic transitions.

One way to assess the effectiveness of the segmentation procedure is to search through the multi-level dendrogram for a path that best matches the time-aligned phonetic transcription. As an example, the best matching path is highlighted in white in Figure 4.1. In the alignment between the dendrogram boundaries with the phone boundaries in the transcription, three kinds of errors can arise. In the first case, the acoustic region can be mapped into a phone, with some time differences between corresponding boundaries. The second case is the deletion of a phonetic boundary in the dendrogram path, and the third is the insertion of an extra acoustic boundary in the dendrogram path. To search for the best path in the dendrogram [11], each possible pathway is scored with the sum of these three kinds of error. The best matching pathway is defined as the one yielding minimum error.

4.3 Distance Metrics

As was mentioned in the previous section, the algorithm for dendrogram segmentation utilizes two distance metrics - the *association distance* for generating

the seed regions, and the *region distance* required by the merging procedure. In our experiments, the association and region distance metrics are kept the same, and the Euclidean distance is used. Since it has been noted that the Euclidean metric over-emphasizes the total gain in the region, and minimizes the importance of spectral shape [39], a *normalized* Euclidean distance has also been included in our experiments. Specifically, the Euclidean distance between two vectors $\vec{x}$ and $\vec{y}$, is divided by the normalized dot product,

$$NormalizationFactor = \frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\|\|\vec{y}\|}$$

It is easy to visualize that this normalizing factor is close to one if $\vec{x}$ and $\vec{y}$ have very similar spectral shapes, but much smaller if the spectral shapes are very different. In the former case, the resulting distance is essentially the same as the Euclidean distance, but in the latter case, the resulting distance is magnified significantly.

4.4 Description of Experiment

Comparison is based on segmenting 500 utterances from 100 speakers of the TIMIT corpus. These sentences contain 19,155 phones. For each acoustic representation, two dendrograms are constructed for every sentence - one uses the Euclidean distance and the other employs the normalized Euclidean metric. The insertion and deletion are then tabulated individually. The overall results are summarized in Figure 4.2.

4.5 Discussion

The mean rate response with normalized Euclidean distance and the MFSC with Euclidean distance performed comparably well with dendrogram segmentation, and better than the DFT. Normalizing the distance metric did not have much effect on the DFT, and yet it increased the amount of insertions of the MFSC (from 5.3% to 8.3%), and reduced both insertion and deletion rates of

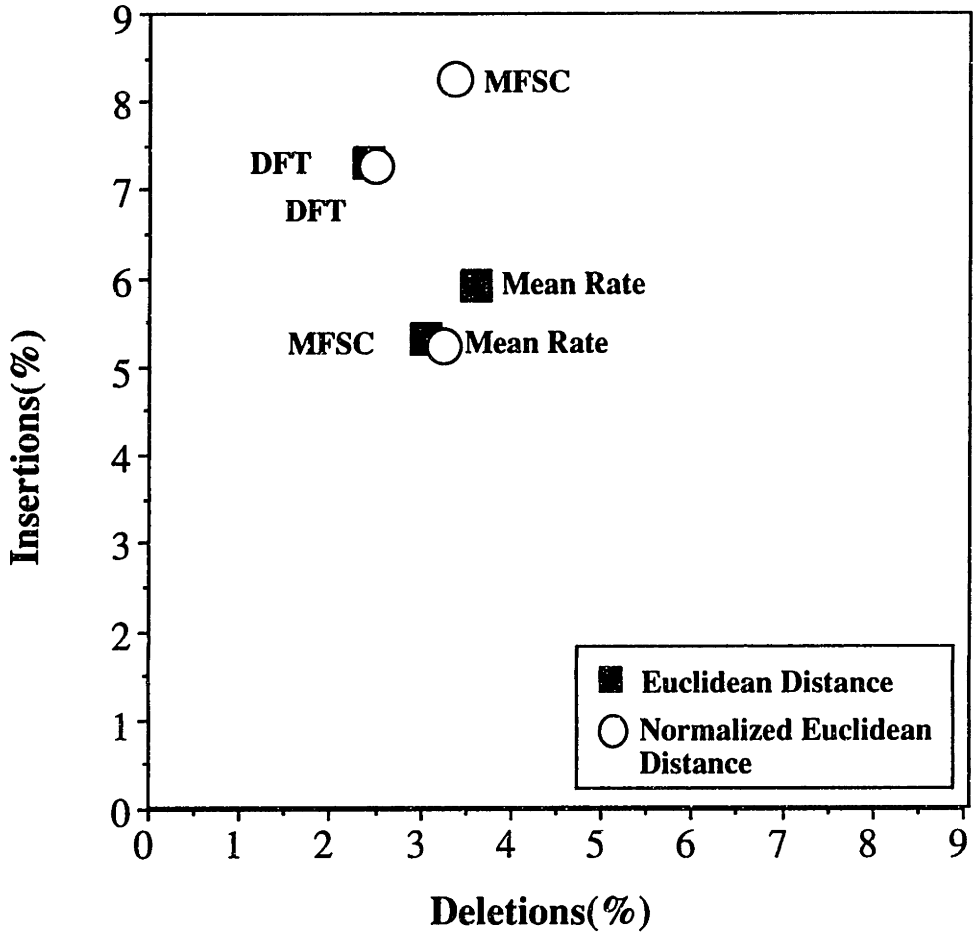

Figure 4.2: Insertion and deletion errors in dendrogram segmentation using three different acoustic representations

the mean rate response (by 0.7% and 0.3% respectively). Closer examination of the magnitudes of the spectral vectors sheds some light on these discrepancies. The mean rate coefficients all lie within the range of 0 to 7, and the normalized Euclidean distance works quite well at capturing the effect of spectral shape similarity. The magnitudes of the DFT are relatively much larger (mostly well below -100 dB), which means that the normalization factor, or cosine of the angle between $\vec{x}$ and $\vec{y}$, tends to be close to 1 regardless of spectral shape similarities. The normalized Euclidean distance does not work well for the MFSC at all because the MFSC coefficients typically varies between -40 dB and 60 dB, which complicates the normalization factor with a sign change. It is deduced that for the sake of comparison, a better-suited normalization factor for the Euclidean distance metric may be [12]:

$$Normalization\ Factor = \frac{1}{2} * (1 + \epsilon + \frac{(\vec{x} - \bar{x}) \cdot (\vec{y} - \bar{y})}{\|\vec{x} - \bar{x}\|\|\vec{y} - \bar{y}\|})$$

where $\bar{x}$ and $\bar{y}$ denote the mean of $\vec{x}$ and $\vec{y}$ respectively, and ϵ is a small additive constant to ensure that the normalization factor is positive.

This normalization factor should range from 0 to 1 with increasing similarity in the spectral shapes between $\vec{x}$ and $\vec{y}$.

Based on our present results, we may tentatively conclude that the MFSC and the mean rate response perform equally well, and they both performed better than the DFT. However, the results are highly dependent on the distance metric used.

4.6 Chapter Summary

In this chapter, we have reported preliminary experimental results on the comparison of three acoustic representations for dendrogram segmentation - the mean rate response from Seneff's auditory model, the MFSC and the DFT. The insertion and deletion rates are tabulated in each case and it is found that the mean rate with a normalized Euclidean distance metric and the MFSC

with a simple Euclidean distance metric have performed comparably well and better than the DFT.

Chapter 5

Conclusions and Future Work

5.1 Summary and Conclusions

In this thesis, we have made an initial attempt to assess the usefulness of distinctive features for phonetic recognition. Distinctive features are a compact inventory of linguistically-motivated units which can be used to concisely describe the many variations in speech such as contextual and coarticulatory phenomena. Therefore, they can potentially serve as powerful data reduction and refinement schemes in tasks of automatic speech recognition, where problems such as variations in speech and sparse training data prevail.

In order to exploit the advantages of distinctive features in the task of speech decoding, we need to first determine how these features are related to the speech signal. Distinctive features manifest themselves as their acoustic correlates in the speech signal, but the nature and characteristics of these acoustic correlates, as well as how they can be captured in the speech signal, are not well understood. In an attempt to extract acoustic attributes which bear some information on the acoustic correlates, it is crucial to select an appropriate acoustic representation. This will involve comparing acoustic representations and choosing the best one. The procedure by which this comparison can be done, however, is not clear. One can start with a set of defined attributes and then decide which acoustic representation will give the best feature extraction results. Alternatively, one can begin with an acoustic

representation believed to be superior to other representations, and then attempt to define and quantify some acoustic attributes. There is no apparent reason for choosing one of these two approaches over the other. In this thesis, the latter approach is adopted.

Chapter 2 describes a comparative study of acoustic representations for vowel classification using the multi-layer perceptron. The representations compared include those originating from Seneff's auditory model, the mel-frequency representations and the Discrete Fourier Transform. The combined outputs of Seneff's auditory model (SAM PC) gave the highest classification performance with both clean and noisy test data. The next best representation was the mean rate response, followed by the synchronous response, the mel-frequency representations (MFSC and MFCC) and the Discrete Fourier Transform (DFT), in that order. Under the assumption that the acoustic representation yielding the best vowel classification performance should be most appropriate to be used for characterizing and quantifying the distinctive features for vowels, we should select SAM PC for our further experiments. However, SAM PC is heterogeneous in nature since half of the representation corresponds to a rotated synchrony spectrum and the other half a rotated mean rate spectrum, attribute extraction can be more conveniently done using a spectral representation. Therefore, the mean rate response, which was the second best representation, was chosen for our further studies.

Chapter 3 describes different methods of incorporating distinctive features into the speech decoding framework. Acoustic attributes which are feature-based have been used in place of the spectral representation. In addition, attempts have been made to map the acoustics into an intermediate phonological representation of distinctive features, which are in turn combined to yield vowels either by table lookup or feature classification. In other words, the overall experiment compares six vowel classification methods, which result from varying three conditional variables, namely, whether acoustic attributes are extracted, whether an intermediate phonological representation was in-

troduced, and whether feature classification is performed. The measurements used as acoustic attributes are based on the spectral center of gravity, the frequency edges of which are optimized for feature distinction in vowels. Our experimental results show that attribute extraction serves as a useful data reduction and refinement scheme. It can reduce the input dimensions approximately by a factor of two, and bring about subsequent computational savings by the same proportion, without any significant loss in vowel classification performance. We are also able to implement the intermediate phonological representation of features without significant deterioration in performance.

The thesis has focused on the task of vowel classification, where the left and right boundaries of a vowel token have been given through a hand-labelled procedure. In order to automate this process, the problem of segmentation of speech is important. Chapter 4 addresses this issue by comparing the relative merits of three acoustic representations in dendrogram segmentation. Dendrogram segmentation aims at constructing a multi-level representation that enables us to capture gradual and abrupt changes through a single hierarchical structure. The acoustic representations studied include the mean rate response, the MFSC and the DFT. We found that using a different acoustic representation demands adopting a suitable distance metric, and therefore fair comparison is not easy to achieve. Nevertheless, the mean rate and the MFSC seem to work comparably well, and better than the DFT, within the context of our experiments.

5.2 Future Work

5.2.1 Improvement with an Intermediate Representation

The paradigm that we have been exploring involves an intermediate representation of acoustic units (acoustic attributes) or/and phonological units (distinctive features) between the signal and the lexicon, as opposed to the direct

classification of phonemes from the signal which has no intermediate representation at all. These two experimental pathways are illustrated in Figure 5.1.

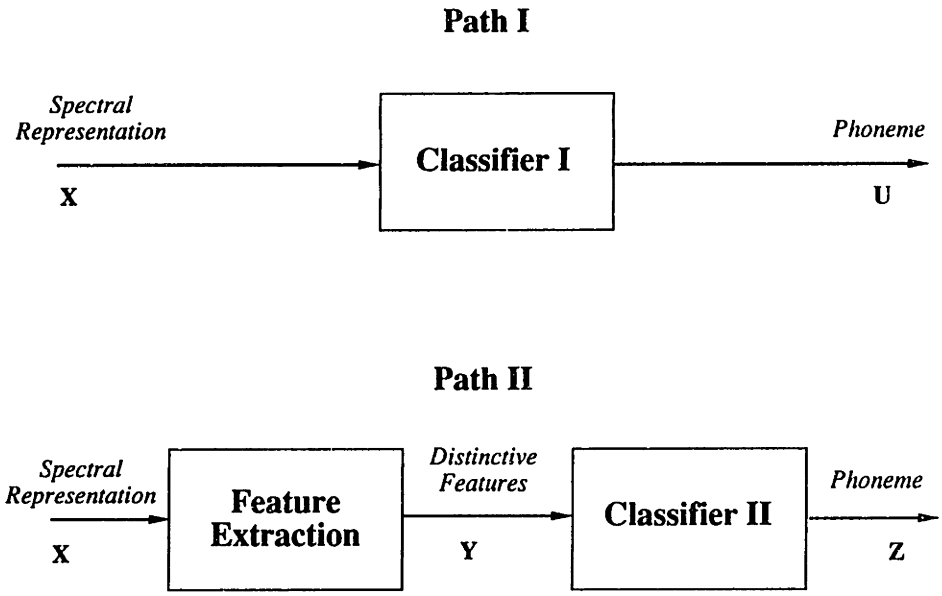

Figure 5.1: Phoneme classification with and without an intermediate representation.

From an information theoretic point of view, where we assume that all probabilities are known, or that infinite training data is available, information is lost as more processing is done. This notion is captured by the Data Processing Theorem [8]. Referring to path II in Figure 5.1, the theorem states that:

If x, y and z form a "Markov Sequence", i.e. the processed output z depends on x only through y, or $Pr(z|x,y) = Pr(z|y)$, for all z and all possible x and y, then

$$I(X;Z) \leq I(X;Y)$$

Proof:

a three-layer MLP (as in path A) for direct vowel classification, but it is conceivable that the additional layers may be more capable of generalizing the acoustics. Therefore it is not certain whether the two-step classification process is inherently sub-optimal to the single-step classification process.

The data processing theorem, however, does shed some light on the following aspect. If mapping the acoustic signal into an intermediate representation (path II of Figure 5.1) is a less severe decision procedure than direct vowel classification (path I), i.e. $I(X;Y) \geq I(X;U)$, and if further processing introduces information loss, then it may perhaps be advantageous to bypass the feature classification process through representing the lexicon in terms of distinctive features. This, of course, will lead to a whole new series of problems which are beyond the scope of this thesis.

The following is a sketch of several directions in which the work in this thesis can be extended. The suggestions, however, are by no means exhaustive.

5.2.2 Extracting Acoustic Attributes

The spectral center of gravity measurement used in our Chapter 3 experiments was chosen because it seems to be a reasonable attribute which carries some kind of formant information. There is certainly room for improvement here. Duration of a vowel token can be included, since it is a good acoustic correlate for the feature [TENSE]. The feature [ROUND] tends to lower the second formant of a vowel towards the first formant, which results in a prominent spectral peak in the low frequency region. Another example which is more applicable to consonants is the concentration of energy within certain frequency bands, such as the concentration of frication energy above 4kHz for alveolar fricatives like /s/ and /z/, as opposed to palatals like /š/ and /ž/, whose frication energy goes well below the 4kHz cutoff. Aside from looking at different frequency bands, dynamics in the time domain are also important. The diphthongs are highly characterized by their longer duration and formant movement - the upward movement of the second formant from [+BACK] to [-BACK] as in /a^y/

and /e^{y}/, or contrarily, the downward movement of the second formant from [+BACK] to [+ROUND] as in /a^{w}/.

Besides, in the extraction of acoustic attributes based on distinctive features, it is important that we distinguish using features that are *produced* from using features that are *intended.* This problem mainly stems from the effects of contextual variation and coarticulation. These effects may exist to the extent that the identity of the phoneme is changed. For example, the vowel in "dwell" is probably intended to be an /ɛ/ which is [-HIGH], [-LOW] and [-BACK], but is often produced like an /ʌ/ which is [-HIGH], [-LOW] and [+BACK], and sometimes even /ɪ/ which is [+HIGH] and [-BACK]. Due to the existence of such discrepancies, special care is required in the association of extracted attributes with certain features. Another example is provided by the [BACK] vowel /u/ which is fronted to from /ü/ in alveolar contexts as in "Tuesday". In this case, attention should be paid to the context and the vowel /ü/ should be trained as [-BACK] rather than [+BACK]. We may also weigh [-HIGH] and [-LOW] as more likely to be produced than [+BACK] in the identification of /u/ or /ü/. It is believed that using features that are *produced* is more advantageous than using those that are *intended*, simply because under many circumstances, the latter cannot be objectively defined.

5.2.3 Feature Classification

We have already seen from our experiments in Chapter 3 that in the process of combining features to give the vowel, using a table lookup procedure led to a significant performance deterioration, but performing feature classification by an MLP does not. The table lookup procedure puts equal emphasis on all features, but the feature classfication procedure does not. This implies that weights should be assigned to individual features for vowel or phoneme classification. It also seems that this set of weights should vary from vowel to vowel, or more generally, from phoneme to phoneme. For example, the features [+HIGH], [+BACK], [+ROUND] and [+TENSE] tend to be more important for

characterizing the vowel /u/ than the feature [RETROFLEX]. On the other hand, the feature [RETROFLEX] is more indicative of the vowel /ɝ/ than other features.

There is also some evidence that the distinctive features show hierarchy in their structure [36]. Manner features, such as [NASAL], [VOICE], [STRIDENT], may be more "fundamental" or higher in the hierarchy than place features since it is possible to identify them reliably simply by observing the speech waveform and without utilizing any contextual information. For example, in consonants - /t/, /d/, /s/, /z/ and /n/ are all [ALVEOLAR], which has, as its acoustic correlates, a second formant centered just below 2 kHz, and major concentration of energy above 4kHz in frication or burst. If we were to determine whether a consonant is [CORONAL], and this consonant neighbors an unstressed vowel, the formant transitions in the vowel cannot be used as a robust cue. But before we start searching for the energy concentration in the frication or the burst, we would need to identify whether the consonant is [NASAL], because the consonant /n/ has neither a burst nor frication. This example serves to illustrate that there is some sort of hierarchy in the features concerning consonants. As for vowels, it seems that the features [HIGH], [LOW] and [BACK] can be more readily identified from the vowel formants and are therefore considered as more fundamental than [ROUND] and [TENSE]. [-BACK] vowels are never [+ROUND], and among the [+BACK] vowels, [ROUND] and [TENSE] are not distinctive, since knowledge of these two features cannot enable us to resolve between /a/ and /ʌ/. The feature [RETROFLEX] is unique because its acoustic correlate - lowering of the third formant - is quite robust even with contextual variations. Our feature mapping experiments have also yielded the highest accuracy for this feature. Furthermore, [RETROFLEX] in a vowel strongly indicates that the vowel is [-HIGH] and [- LOW]. This redundancy between features can probably be exploited in attribute extraction, since for some features like [TENSE], [+FEATURE] may be easier to detect in the acoustic signal than [-FEATURE], or vice versa.

The geometry of distinctive features has yet to be determined. Performance improvement may perhaps be achieved by reliable extraction of the features high in the hierarchy, or the features that are more likely to be produced than others within a particular context.

5.2.4 Acoustic Segmentation

The preliminary experiments on dendrogram segmentation have shown that the ability to capture acoustic landmarks is sensitive to the choice of the distance metric for different acoustic representations. A better distance metric is required in order to conduct a fair comparison among different acoustic representations. In addition, it may also be interesting to find out whether certain acoustic representations can preserve acoustic regularities better than others in the presence of noise.

Appendix A

The Mel-frequency Cosine Transform

This appendix attempts to provide a brief explanation of the cosine transform employed in obtaining the mel-frequency cepstral coefficients (MFCC) from the mel-frequency spectral coefficients (MFSC). The idea of the computation is to treat the MFSC as the Discrete Fourier Transform of the MFCC.

As has been mentioned in Chapter 2, the MFSC are obtained from performing bandpass summation on the power spectrum of a windowed speech signal through a series of 40 overlapping triangular filters. The log energy output of each filter together form the MFSC - denoted by $X_k, k = 1, 2, 3...40$. In order to treat this as the Discrete Fourier Transform of a real speech signal, we have to impose even symmetry by folding the spectrum about an edge, as illustrated in Figure A.1 [1]

Therefore, we can see that $X_1 = X_0$, $X_2 = X_{-1}$, $X_3 = X_{-2}$, etc., and the symmetry lies about the axis of $k = \frac{1}{2}$. In other words, if we shift our reference origin to $k = \frac{1}{2}$, our spectrum becomes even symmetric in that $X_{0.5} = X_{-0.5}$, $X_{1.5} = X_{-1.5}$, $X_{2.5} = X_{-2.5}$, etc., and we will be able to obtain a real signal by performing an 80-point Inverse Discrete Fourier Transform (IDFT).

[1] This method of imposing even symmetry will preserve the maximum number of degrees of freedom when one does the inverse transform, i.e. in this case, we can use an 80-point IDFT. But if, for example, we re-label the indices for k to range between 0 to 39, the even symmetry so created corresponds to a 78-point IDFT.

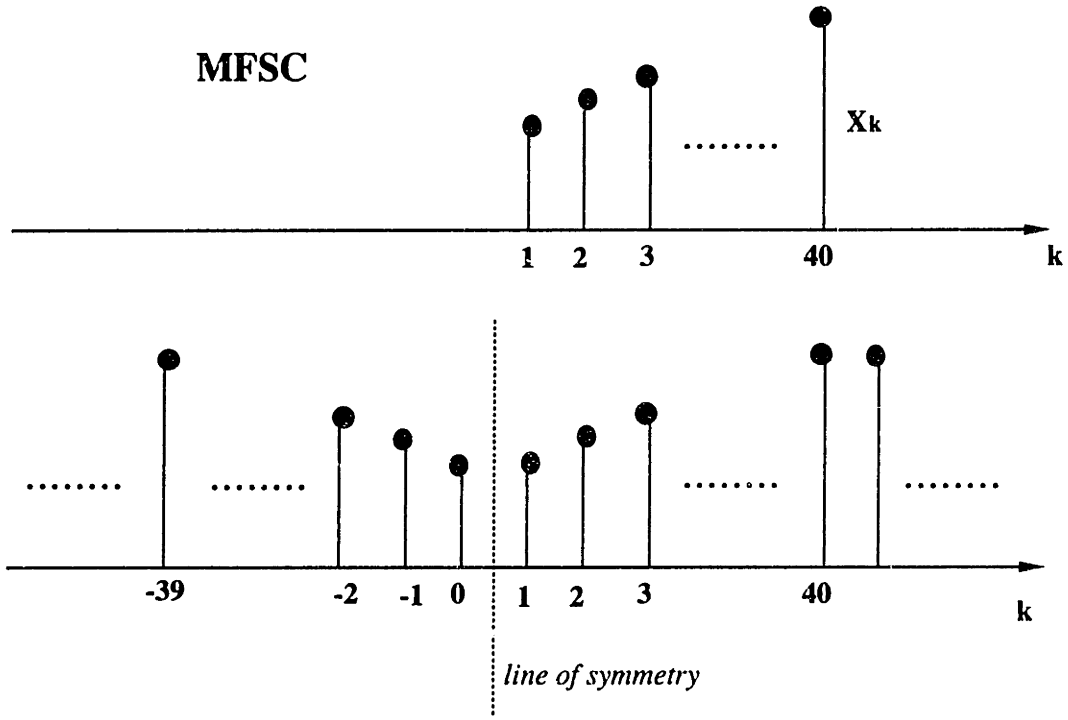

Figure A.1: Imposing even symmetry on the spectrum of MFSC by folding about an edge

The IDFT equation for an 80-point DFT is:

$$x[n] = \sum_{k'=0}^{k'=79} X_k e^{j\frac{2\pi}{80}k'n} \tag{A.1}$$

Shifting our reference origin to $k = \frac{1}{2}$, we have

$$\begin{aligned} x[n] &= \sum_{k'=0.5}^{k'=79.5} X_{k'} e^{j\frac{2\pi}{80}k'n} \\ &= \sum_{k'=0.5}^{k'=39.5} X_{k'} e^{j\frac{2\pi}{80}k'n} + \sum_{k'=40.5}^{k'=79.5} X_{k'} e^{j\frac{2\pi}{80}k'n} \\ &= \sum_{k'=0.5}^{k'=39.5} X_{k'} e^{j\frac{2\pi}{80}k'n} + \sum_{k'=-39.5}^{k'=-0.5} X_{k'} e^{j\frac{2\pi}{80}k'n} \\ &= \sum_{k'=0.5}^{k'=39.5} X_{k'} e^{j\frac{2\pi}{80}k'n} + \sum_{k'=0.5}^{k'=39.5} X_{k'} e^{-j\frac{2\pi}{80}k'n} \\ &= \sum_{k'=0.5}^{k'=39.5} X_{k'} cos(\frac{\pi}{40}k'n) \end{aligned} \tag{A.2}$$

where we have the property of even symmetry. Finally, recall that k ranges from 1 to 40 whereas k' ranges from 0.5 to 39.5, so substituting variables, we obtain:

$$x[n] = \sum_{k=1}^{k=40} X_k cos[\frac{\pi}{40}(k - \frac{1}{2})n] \tag{A.3}$$

which is our cosine transform equation.

Appendix B

Detailed Statistics on Relative Vowel Classification Performances

As was mentioned in Chapter 2, the six acoustic representations are compared by conducting vowel classification experiments. There are a total of four separate experiments where the number of training tokens is increased from 2,000, to 4,000, 8,000 and finally 20,000. Figure B.1 displays the average test set accuracies over six iterations of each experiment. The fluctuation in performance between successive iterations lie around 1%.

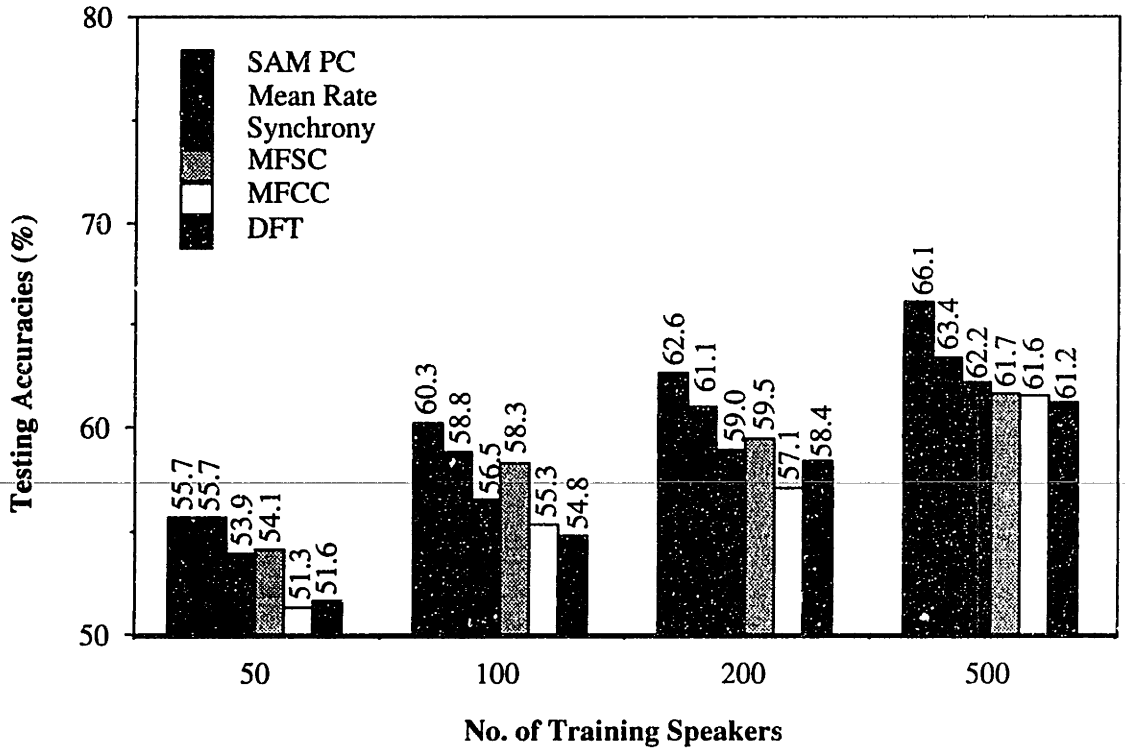

Figure B.1: Overall comparison results for the six acoustic representations

Appendix C

Acoustic Attributes

This appendix lists the set of acoustic attributes used in the experiments reported in Chapter 3. An acoustic attribute for a given feature includes the frequency and amplitude of a spectral center of gravity, which is computed between an optimized pair of lower and upper frequency edges. The pair of free parameters is individually optimized for each third of the vowel token, so as to implicitly capture the dynamics of articulation. In the following tables, a frequency edge is expressed as a coefficient index in the mean rate response. Since these coefficients are spaced a half-Bark apart, dividing the coefficient index by two gives the correspond frequency in Barks.

The features [HIGH] and [LOW] use two attributes each. In Table C.1, the first row displays the set of free parameters giving the maximum separation between the classes [+HIGH] and [-HIGH] measured with the FDC score, and the second row displays the free parameters giving the second highest FDC score. Similarly, the feature [LOW] also has two sets of free parameters. The remaining features - [BACK], [TENSE], [ROUND] and [RETROFLEX] use only one attribute per feature.

	Initial third of token		*Middle third of token*		*Final third of token*	
	lower edge	*upper edge*	*lower edge*	*upper edge*	*lower edge*	*upper edge*
High	0	14	0	16	0	17
High	8	33	8	32	9	34
Low	0	14	0	14	0	15
Low	8	30	9	32	8	34
Back	12	30	13	29	12	30
Tense	0	17	0	20	0	21
Round	0	36	0	36	12	30
Retroflex	19	34	19	34	19	37

Table C.1: Acoustic attributes with optimized free parameters

Appendix D

Confusion Matices

The following are the confusion matrices obtained from experiments conducted under conditions A through F in Chapter 3.

	e	æ	i	ɛ	ɪ	o	u	a	ɔ	ɝ	ʌ	ü	ʊ	Total
e	65	4	7	8	11	0	0	1	0	3	1	0	0	135
æ	4	63	0	15	7	1	0	6	2	0	3	0	0	137
i	5	1	87	0	4	0	0	0	0	0	0	2	0	266
ɛ	8	10	1	46	16	2	0	1	1	7	7	0	0	158
ɪ	7	4	9	10	61	0	1	0	0	2	1	2	2	218
o	3	3	0	6	1	61	3	1	8	2	12	0	0	99
u	0	0	7	4	11	4	44	0	4	7	0	19	0	27
a	0	3	0	1	0	2	0	74	12	1	7	0	0	167
ɔ	0	0	0	0	1	11	0	22	63	1	1	0	1	140
ɝ	1	1	0	0	2	0	0	1	0	91	0	2	0	82
ʌ	0	9	0	10	7	2	1	14	4	3	48	0	2	128
ü	1	2	14	1	8	0	11	0	0	2	0	59	0	83
ʊ	0	3	0	0	38	13	6	0	9	0	16	0	16	32
Total	140	142	277	149	230	91	29	186	130	109	108	67	14	

Table D.1: Confusion Matrix for Condition A - Classification of the Mean-Rate Response into Vowels

	e	æ	i	ɛ	ɪ	o	u	a	ɔ	ɝ	ʌ	ü	ʊ	Total
e	65	2	8	7	16	0	0	1	0	0	1	0	0	135
æ	1	73	0	10	6	0	0	3	0	1	6	0	0	137
i	5	0	85	0	6	0	0	0	0	0	0	3	0	266
ɛ	8	16	2	44	18	1	0	2	0	4	5	0	0	158
ɪ	4	4	7	8	69	0	0	0	0	3	2	1	0	218
o	2	2	0	5	3	55	5	1	10	3	13	0	1	99
u	0	0	0	0	15	7	41	0	7	7	4	19	0	27
a	0	2	0	2	0	1	0	77	8	1	8	0	0	167
ɔ	1	0	0	1	0	12	0	30	53	1	2	0	1	140
ɝ	0	1	0	0	5	0	0	1	0	85	1	6	0	82
ʌ	0	9	0	20	5	1	0	16	3	3	43	0	0	128
ü	1	2	16	0	14	2	10	0	0	8	0	46	0	83
ʊ	0	0	0	6	38	13	3	0	9	6	22	3	0	32
Total	129	158	269	147	267	84	26	203	107	104	117	58	3	

Table D.2: Confusion Matrix for Condition B - Classification of Attributes into Vowels

	e	æ	i	ɛ	ɪ	o	u	a	ɔ	ɝ	ʌ	ü	ʊ	Total
e	71	2	7	4	13	0	0	0	0	0	1	0	0	135
æ	2	72	1	16	4	0	0	3	0	1	2	0	0	137
i	7	0	85	0	6	0	0	0	0	0	0	2	0	266
ɛ	10	14	1	44	17	1	0	1	0	4	8	0	0	158
ɪ	7	4	9	8	63	1	0	0	0	4	1	2	1	218
o	2	1	0	3	2	62	2	2	10	4	11	0	0	99
u	0	0	0	0	4	11	41	0	4	4	4	26	7	27
a	0	2	0	1	0	2	0	77	10	1	7	0	0	167
ɔ	1	0	0	0	1	13	0	26	55	1	3	0	1	140
ɝ	0	0	0	2	4	0	1	1	0	87	1	4	0	82
ʌ	0	9	0	16	7	2	0	15	2	2	47	0	1	128
ü	1	0	22	0	14	2	12	0	0	5	0	42	1	83
ʊ	0	3	0	6	38	9	6	0	6	6	19	3	3	32
Total	154	149	274	143	243	99	26	190	110	103	116	55	10	

Table D.3: Confusion Matrix for Condition D - Classification of Attributes into Features and then to Vowels

	e	æ	i	ɛ	ɪ	o	u	a	ɔ	ɝ	ʌ	ü	ʊ	Total
e	67	4	6	7	15	0	0	0	0	0	1	0	0	135
æ	2	72	1	15	4	1	0	2	0	0	3	0	0	137
i	7	0	84	0	6	0	0	0	0	0	0	2	0	266
ɛ	8	13	1	41	23	0	0	1	0	3	10	0	0	158
ɪ	10	4	8	8	64	0	0	0	0	3	1	3	0	218
o	2	3	0	3	1	56	3	2	11	4	14	1	0	99
u	0	0	0	0	11	11	44	0	4	4	4	19	4	27
a	0	3	0	1	0	1	0	69	14	1	10	0	0	167
ɔ	0	1	0	0	0	11	0	22	62	2	2	0	0	140
ɝ	1	1	0	4	2	0	1	1	0	82	2	5	0	82
ʌ	0	9	0	16	9	4	0	13	2	5	38	1	2	128
ü	1	1	19	1	6	0	14	0	0	2	0	54	0	83
ʊ	0	3	0	3	44	13	6	0	6	3	16	0	6	32
Total	149	159	267	143	255	85	31	171	128	98	114	67	5	

Table D.4: Confusion Matrix for Condition F - Classification of the Mean-Rate Response into Features and then into Vowels

	e	æ	i	ɛ	ɪ	o	u	a	ɔ	ɝ	ʌ	ü	ʊ	Total
e	57	0	15	19	8	0	0	0	0	0	1	0	0	135
æ	0	45	1	40	7	0	0	3	0	0	3	0	0	137
i	5	0	89	0	5	0	0	0	0	0	0	1	0	266
ɛ	5	9	3	58	11	0	0	1	0	0	9	0	0	158
ɪ	2	1	15	17	53	0	0	0	0	1	1	1	1	218
o	1	1	0	7	0	52	2	4	4	0	16	0	0	99
u	0	0	4	0	0	7	41	0	0	0	4	22	7	27
a	0	1	0	1	0	1	0	75	10	0	9	0	0	167
ɔ	1	0	0	0	0	9	1	27	46	0	6	0	0	140
ɝ	0	0	1	2	1	0	0	1	0	70	7	1	2	82
ʌ	1	6	0	23	5	0	0	15	2	2	43	0	0	128
ü	1	0	29	4	8	1	12	0	0	2	0	34	1	83
ʊ	0	0	0	13	34	6	9	0	3	0	13	3	0	32
Total	106	91	320	256	190	70	27	192	88	63	129	41	8	

Table D.5: Confusion Matrix for Path C - Classification of Attributes into Features followed by table-lookup

	e	æ	i	ɛ	ɪ	o	u	a	ɔ	ɝ	ʌ	ü	ʊ	Total
e	61	2	10	17	7	0	0	0	0	0	1	0	0	135
æ	1	52	1	32	1	1	0	7	0	0	3	0	0	137
i	6	0	88	1	3	0	0	0	0	0	0	2	0	266
ɛ	7	8	2	54	11	0	0	1	0	1	10	0	0	158
ɪ	6	2	13	17	49	0	0	0	0	2	2	3	0	218
o	1	1	0	5	0	48	2	2	6	0	14	1	1	99
u	0	0	0	0	11	11	44	0	4	4	0	19	0	27
a	0	2	0	1	0	1	0	75	14	1	5	0	0	167
ɔ	0	0	0	0	0	8	0	24	56	0	2	0	0	140
ɝ	1	1	0	4	1	0	1	0	0	74	4	4	0	82
ʌ	0	5	0	17	5	0	0	16	3	3	38	1	1	128
ü	0	1	23	1	2	0	12	0	0	0	0	53	1	83
ʊ	0	3	0	3	34	9	6	0	6	0	12	0	3	32
Total	125	106	300	227	165	67	28	193	115	74	106	65	3	

Table D.6: Confusion Matrix for Path E - Classification of the Mean Rate Response into Features followed by table-lookup

Bibliography

[1] Allen, J. B., "Cochlear Modelling", *IEEE ASSP Magazine*, vol. 2, no. 1, pp.3-29, 1985.

[2] Chomsky N. and M. Halle, *Sound Pattern of English*, Harper & Row, 1968

[3] Cohen, J. R. "Application of an adaptive auditory model to speech recognition", *Proc. Symp. on Speech Recognition*, Montreal, pp.8-9, July, 1986.

[4] Dillon, W. R.and M. Goldstein, "Multivariate Analysis methods and applications", *Wiley Series in Probability and Mathematical Statistics.*

[5] Duda, R.O. and P.E. Hart, "Pattern Classification and Scene Analysis", a Wiley-Interscience publicaton, 1973.

[6] Fant, G., *Analysis and Synthesis of Speech Processes*, Chapter 8, *Manual of Phonetics*, Bertil Malmberg, ed. North Holland Press, 1970.

[7] Fant, G., *Acoustic Theory of Speech Production*, Mouton and Co., 's-Gravenhage, Netherlands, 1960.

[8] Gallager, R. G., *Information theory and reliable communication*, John Wiley and Sons, 1968.

[9] Gillick, L. and S. J. Cox, "Some statistical issues in the comparison of speech recognition algorithms," *ICASSP-89*, pp. 532-535, 1989.

[10] Glass, J. R. and V. W. Zue, "Multi-Level Acoustic Segmentation of Contiuous Speech," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, April 1988.

[11] Glass, J. R., "Finding Acoustic Regularities in Speech: Applications to Phonetic Recognition", Ph.D. Thesis, MIT, 1988.

[12] Glass, J. R., personal communication

[13] Hermansky, Hynek, "An Efficient Speaker-independent Automatic Speech Recognition by simulation of some properties of Human Auditory Perception", *Proc. IEEE International Conference of Acoustics, Speech & Signal Processing (ICASSP-87)*, pp.1159-1162, 1987.

[14] Hunt, Melvyn and Lefebvre, Claude, "Speaker Dependent and Independent Speech Recognition Experiments with an Auditory Model", *Proc. ICASSP-88*, New York, pp.215-218, 1988.

[15] Hunt, Melvyn and Lefebvre, Claude, "A Comparison of Several Acoustic Representations for Speech Recognition with Degraded and Undegraded Speech", *Proc. ICASSP-89*, pp.262-265, 1989.

[16] Mermelstein, P. and S. Davis, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", *IEEE Transactions on Acoustics, Speech and Signal Processing*", vol. ASSP-28, No. 4, August 1980.

[17] Itakura, F., "Minimum Prediction Residual Principle Applied to Speech Recognition," *IEEE Trans. Acoustics, Speech, and Signal Proc.*, Vol. ASSP-23, No.1, pp. 67-72, February 1975.

[18] Klatt, D.H., "The problem of variability in speech recognition and in models of speech perception", in *Invariance and Variability in Speech Processes*, J.S. Perkell and D.H. Klatt, Eds. Hillsdale, NJ:Lawrence Erlbaum Assoc., 1985.

[19] Lamel, L. F., R. H.Kassel, and S. Seneff, "Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus," *Proc. DARPA Speech Recognition Workshop,* Report No. SAIC-86/1546, pp. 100–109, February 1986.

[20] Leung, Hong. C. and Zue, Victor. W., "Phonetic Classification Using Multi-Layer Perceptrons", *Proc. ICASSP-90*, Albuquerque, pp.525-528, 1990.

[21] Leung, H.C. *"The Use of Artificial Neural Nets for Phonetic Recognition,"* P.h.D. Thesis, MIT Department of Electrical Engineering and Computer Science, Cambridge, MA, May 1989.

[22] Leung, H.C. and V.W. Zue, Personal communication.

[23] Meng, H.M. and V.W. Zue, " A Comparative Study of Acoustic Representations of Speech for Vowel Classification using Multi-Layer Perceptrons", *Proc. ICSLP-90*, Kobe, pp.1053-1056, 1990.

[24] Meng, H.M. and V.W. Zue, "Signal Representation Comparison for Phonetic Classification," *ICASSP-91*, Toronto, May 1991.

[25] O'Shaughnessy D., *Speech Communication - Human and Machines*, Addison-Wesley Publishing Company, 1987.

[26] Paul, D. P., "A Speaker-stress Resistant HMM Isolated Word Recognition", *Proc. ICASSP-87*, Dallas, pp.713-716, 1987.

[27] Phillips, Michael S., "Speaker Independent Classification of Vowels and Diphthongs in Continuous Speech", *Proc. ICPhS-87*, Tallinn, pp.240-243, 1987.

[28] Rabiner, L. R. and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1978.

[29] Rumelhart, D. E., J. L. McClelland and the PDP Research Group, "Parallel Distributed Processing, MIT Press, 1986.

[30] Seneff, S. "A joint synchrony/mean-rate model of auditory speech processing", *J. Phonetics*, vol. 16, no. 1, pp.55-76, 1988.

[31] Seneff, S., *"Pitch and Spectral Analysis of Speech based on an Auditory Synchrony Model,"* P.h.D. Thesis, MIT Department of Electrical Engineering and Computer Science, Cambridge, MA, May 1985.

[32] Seneff, S., "Vowel recognition based on line-formants derived from an auditory-based spectral representation," *Proc. of the 11th International Congress of Phonetic Sciences*, Estonia, USSR, 1987.

[33] Sorensen, H. B. D. and P. Dalsgaard, "Multi-level Segmentation of Natural Continuous Speech using Different Auditory Front-ends," *Proc. European Conference on Speech Communication and Technology*, September 1989.

[34] Stevens, K. N., R. S. Nickerson and A. M. Rollins, "Suprasegmental and Postural Aspects of Speech Production and their effect on Articulatory Skills and Intelligibility," Chapter 2 in *Speech of the Hearing Impaired*, I. Hochberg, H. Levitt and M. J. Osberger, Eds., University Park Press, Baltimore, MD, 1983.

[35] Stevens, K. N., "The quantal nature of speech: Evidence from articulatory-acoustic data," in *Human Communication: A Unified View*, E. E. David and P. B. Denes, eds., McGraw-Hill, New York, 1972.

[36] Stevens, K., personal communication

[37] Stevens, K.N., Unpublished course notes for *Speech Communications*, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Spring term, 1989.

[38] Zue, V. W., J. R. Glass, M. S. Phillips, and S. Seneff, "Acoustic Segmentation and Phonetic Classification in the SUMMIT System", *Proc. ICASSP-89*, Scotland, pp.389-392, 1989.

[39] Zue, V. W. et al, "Recent Progress on the SUMMIT System," *Proc. the Third DARPA Speech and Natural Language Workshop*, Hidden Valley, June 1990.