Consonant Recognition by Humans and Machines

by

Jason Sroka

Submitted to the Division of Health Sciences and Technology
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 1998

© Massachusetts Institute of Technology 1998. All rights reserved.
Consonant Recognition by Humans and Machines

by

Jason Sroka

Submitted to the Division of Health Sciences and Technology
on September 8, 1998, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

The goal of this research is to determine how aspects of human speech processing can be utilized to improve the performance of Automatic Speech Recognition (ASR) systems. Three traditional ASR parameterizations matched with Hidden Markov Models (HMMs) are compared to humans on a consonant recognition task using Consonant-Vowel-Consonant (CVC) nonsense syllables degraded by highpass filtering, lowpass filtering, or additive noise. Confusion matrices were determined by recognizing the syllables using different ASR front ends, including Mel-Filter Bank (MFB) energies, Mel-Filtered Cepstral Coefficients (MFCCs), and the Ensemble Interval Histogram (EIH).

For syllables degraded by lowpass and highpass filtering, automated systems trained on the degraded condition recognized the consonants roughly as well as humans. Moreover, all the ASR systems produce similar patterns of recognition errors for a given filtering condition. These patterns differ significantly from that characteristic of humans under the same filtering conditions.

For syllables degraded by additive speech-shaped noise, none of the automated systems recognized consonants as well as humans. As with filtered conditions, confusion matrices revealed similar error patterns for all the ASR systems. While the error patterns of humans and machines was more similar for noise conditions than for filtered conditions, the similarities were not as great as between the ASR systems. The greatest difference between human and machine performances was in determining the correct voiced/unvoiced classification of consonants.

Given these results, work was focused on recognition of the correct voicing classification in additive noise (0 dB SNR). The approach taken attempted to automatically extract attributes of the speech signal, termed subphonetic features, which are useful in determining the distinctive feature voicing.

Two subphonetic features, intervocalic period (the length of time between the onset of the vowel and any preceding vocalization) and delta fundamental (the average first difference of fundamental frequency over the first 90 msec of the vowel) proved particularly useful. When these two features were appended to traditional ASR parameters,
the deficit exhibited by automated systems was reduced substantially, though not eliminated.

Thesis Supervisor: Louis D. Braida
Title: Henry E. Warren Professor of Electrical Engineering
Acknowledgments

I have often sat and wondered at the meaning of it all; at what single idea could explain the world in all its beautiful detail or what single element pervaded everything around me. Naturally, these thoughts, while not fruitless, never yielded an answer. This research takes a different approach to understanding, reducing the scope of contemplation to pursue a much more well-defined question. It is up to the future to determine whether a synthesis of the parts can ever explain or enlighten on the whole. I take pleasure in my contemplation of the whole and pursuit of the part.

The work embodied in this thesis would never have been possible without the efforts of a great many people. The thesis committee, Paul Duchnowski, Bertrand Delgutte, Rich Lippmann, Ken Stevens, and especially my advisor Lou Braida, helped greatly in defining and pursuing the goals of the research.

Many thanks go to everyone in the Sensory Communications Group, unfortunately too numerous to name, who provided a great environment in which to pursue this research. Their friendliness and eagerness to help made the day-to-day efforts involved in this research enjoyable.

I would like to thank the members of the auditory journal group, Rob Baxter, Ben Gold, and Tom Quatieri. The journal group kept my intellectual curiosity alive when I was doing little else beside this thesis research, and helped me greatly in developing my thesis defense.

Finally, I would like to acknowledge a huge debt of gratitude to the Speech and Hearing Sciences program, the admissions committee that accepted my application, and especially Nelson Kiang, for providing me with the opportunity to pursue a childhood dream. I have found the professors and the other students in the Speech and Hearing Sciences Program to be of a uniformly high caliber, and will always consider myself fortunate to have benefited from their tutelage and friendship.
Contents

0.1 Abstract ......................................................... 13

1 Introduction .................................................. 15
   1.1 Automated Speech Recognition ......................... 16
   1.2 Using Knowledge of Human Audition for Automated Speech Recognition 17
      1.2.1 Auditory-Based Front Ends .......................... 18
      1.2.2 Auditory-Based Back Ends ......................... 21
   1.3 Overview of the Thesis .................................. 22

2 The Nonsense Syllable Databases .......................... 23

3 The Automated Recognition Systems ...................... 26
   3.1 The MFB and MFCC Parameterizations .................. 26
   3.2 Ensemble Interval Histogram (EIH) .................... 27
   3.3 Training Approaches .................................. 30

4 Comparisons of Human and Machine Consonant Recognition 32
   4.1 Error Rate Comparisons ................................ 32
      4.1.1 Highpass Filtering Results ....................... 32
      4.1.2 Lowpass Filtering Results ....................... 35
      4.1.3 Additive Noise .................................. 36
   4.2 Information Transfer Analysis of Confusion Matrices ... 40
      4.2.1 Highpass Filtering Results ....................... 41
      4.2.2 Lowpass Filtering Results ....................... 43
4.2.3 Additive Noise Results ........................................... 43

4.3 Comparison of Consonant Pair Confusions ......................... 46

4.3.1 Metric Multidimensional Scaling (MMS) .......................... 46

4.3.2 The Scaling Model .............................................. 48

4.3.3 MMS Analysis of Phone Confusion Matrices ....................... 49

4.4 Summary of Comparison Results .................................... 57

5 Subphonetic Features Used by Humans for Voicing Classification 60

5.1 Plosive Voicing Cues ............................................. 61

5.2 Fricative Voicing Cues ........................................... 68

6 Subphonetic Feature Extraction ...................................... 73

6.1 Waveform Analysis Methods ....................................... 73

6.1.1 Spectrogram Analysis ........................................... 73

6.1.2 Short-Time Autocorrelation Analysis ............................ 80

6.2 Subphonetic Feature Detection ..................................... 82

6.2.1 Landmark Detection ............................................ 83

6.2.2 Intervocalic Period ............................................. 83

6.2.3 Voice Onset Time .............................................. 86

6.2.4 Delta Fundamental ............................................ 90

7 Subphonetic Feature Extraction Results .............................. 92

7.1 Voicing Classification Scores Using Subphonetic Features ........ 92

7.2 Use of IVP and Delta Fundamental in Quiet ........................ 94

7.3 Orthogonality of Subphonetic Features and Traditional Systems . 94

7.4 Integration of Subphonetic Feature Extraction with Traditional Parameterizations ........................................... 95

7.5 Summary of Subphonetic Feature Extraction Results ................ 97

8 Discussion ....................................................................... 100

8.1 Human and Machine Performance Comparison ...................... 100

8.2 Subphonetic Feature Analysis ...................................... 101
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-1</td>
<td>Overview of the Automated Speech Recognition (ASR) process.</td>
<td>16</td>
</tr>
<tr>
<td>3-1</td>
<td>Frequency response for 24 (of 73) filters generated with the Matlab auditory toolkit based on Lyon's passive ear model.</td>
<td>28</td>
</tr>
<tr>
<td>4-1</td>
<td>Percent correct scores for humans and the MFB systems on speech degraded by highpass filtering.</td>
<td>33</td>
</tr>
<tr>
<td>4-2</td>
<td>Percent correct scores for humans and the MFCC systems on speech degraded by highpass filtering.</td>
<td>34</td>
</tr>
<tr>
<td>4-3</td>
<td>Percent correct scores for humans and the EIH systems on speech degraded by highpass filtering.</td>
<td>34</td>
</tr>
<tr>
<td>4-4</td>
<td>Percent correct scores for humans and automated systems using Degradation-Specific Training (DST) on speech degraded by highpass filtering.</td>
<td>35</td>
</tr>
<tr>
<td>4-5</td>
<td>Percent correct from human CV results, predicted CVC scores, and MFB systems for speech degraded by lowpass filtering.</td>
<td>36</td>
</tr>
<tr>
<td>4-6</td>
<td>Percent correct from human CV results and MFCC systems for speech degraded by lowpass filtering.</td>
<td>37</td>
</tr>
<tr>
<td>4-7</td>
<td>Percent correct from human CV results and EIH systems for speech degraded by lowpass filtering.</td>
<td>37</td>
</tr>
<tr>
<td>4-8</td>
<td>Percent correct scores for humans and the automated systems using Degradation-Specific Training (DST) on speech degraded by lowpass filtering</td>
<td>38</td>
</tr>
<tr>
<td>4-9</td>
<td>Percent correct scores from human data and MFB systems for speech degraded by additive noise.</td>
<td>39</td>
</tr>
</tbody>
</table>
4-10 Percent correct scores from human data and MFCC systems for speech degraded by additive noise. ......................................................... 39
4-11 Percent correct scores from human data and EIH systems for speech degraded by additive noise. ......................................................... 40
4-12 Percent correct scores for humans and the automated systems using Degradation-Specific Training (DST) on speech degraded by lowpass filtering. ................................................................. 41
4-13 Information Transfer (IT) scores from human CVC and CV data and from MFB systems on highpass filtered speech. ......................... 42
4-14 Information Transfer (IT) scores from human CV data and from MFB systems on lowpass-filtered speech. ........................................... 44
4-15 Information Transfer (IT) scores from human CVC and CV data and from MFB systems on speech degraded by additive noise. ............... 45
4-16 d’ scores for MST and DST cepstral systems recognizing speech highpass filtered with a 1600 Hz cutoff frequency. ................................. 48
4-17 d’ scores for MFCC systems recognizing speech degraded by lowpass filtering with a 500 Hz cutoff frequency and additive noise at 0 dB SNR. 53
4-18 d’ scores for MFB and human results for speech in additive noise. ................................................................. 57

5-1 Spectrograms of a voiced (/d/) and unvoiced (/t/) plosive in quiet and in noise. ................................................................. 63
5-2 Spectrograms of the full set of plosives in quiet and noise. ......................... 66

6-1 Illustration of the dual approach to exploring automatic extraction of elements of the speech signal which can be used to distinguish between voiced and unvoiced consonants. ........................................ 74
6-2 An illustration of the spectrographic template approach to analyzing speech. ................................................................. 78
6-3 Example of the short-time autocorrelation smoothness measure. ............ 81
6-4 Example of IVP detection. ................................................................. 85
6-5 Distribution of IVP for unvoiced and voiced initial consonants in 0 dB additive noise. ................................................... 86
6-6 Distribution of VOT values for unvoiced and voiced initial consonants. 89
6-7 Distribution of VOT values for voiced and unvoiced plosives. ........ 90
6-8 Distribution of Delta Fundamental values for unvoiced and voiced ini-
    tial consonants. ...................................................... 91

7-1 Percent correct voicing categorization for the traditional approaches 
    with and without the subphonetic features. ................................. 98

A-1 Information Transfer (IT) scores from human CVC and CV data and 
    from MFCC systems on highpass filtered speech. .......................... 107
A-2 Information Transfer (IT) scores from human CVC and CV data and 
    from MFCC systems on lowpass filtered speech. .......................... 108
A-3 Information Transfer (IT) scores from human CVC and CV data and 
    from MFCC systems on noisy speech. ................................. 109
A-4 Information Transfer (IT) scores from human CVC and CV data and 
    from EIH systems on highpass filtered speech. .......................... 110
A-5 Information Transfer (IT) scores from human CVC and CV data and 
    from EIH systems on lowpass filtered speech. .......................... 111
A-6 Information Transfer (IT) scores from human CVC and CV data and 
    from EIH systems on noisy speech. ................................. 112
List of Tables

2.1 Feature descriptions of the twelve consonants used in the CVC database. 24

3.1 Inter-‘spike’ intervals for each EIH bin. ............................................. 29

4.1 Regression analysis correlation coefficients, standard deviations, and numbers of comparisons for consonant pair comparisons of the automated systems using Multi-Style and Degradation-Specific Training. . 50

4.2 Regression analysis values for consonant pair comparisons of the different front ends. ................................................................. 51

4.3 Regression analysis values for consonant pair comparisons: Initial Consonant Recognition vs Final Consonant Recognition. ............... 51

4.4 Regression analysis values for comparisons of response patterns on the different signal degradations. ............................................. 52

4.5 Regression analysis values for comparisons of response patterns for varying severities of the different types of signal degradation. . ....... 54

4.6 Regression analysis values for consonant pair comparisons between human and machine recognition results. ............................. 56

4.7 Description of the consonant pairs for which human performance in +2 dB additive speech-shaped noise is nearly perfect while the MFB system is making errors in 0 dB additive speech-shaped noise. ......... 58

5.1 List of potential cues to voicing classification of plosives. ............... 71

5.2 List of potential cues to voicing classification of fricatives. ............... 72
7.1 Percent correct voiced/unvoiced categorization based on various combinations of the subphonetic features and comparison values from human and traditional ASR results. .............................. 93

7.2 Percentage of initial consonant tokens which were labelled with different voiced/unvoiced category by pairs of approaches. .............................. 96

7.3 Performance of traditional systems before and after inclusion of sub-phonetic features. .............................. 97
0.1 Abstract

The goal of this research is to determine how aspects of human speech processing can be utilized to improve the performance of Automatic Speech Recognition (ASR) systems. Three traditional ASR parameterizations matched with Hidden Markov Models (HMMs) are compared to humans on a consonant recognition task using Consonant-Vowel-Consonant (CVC) nonsense syllables degraded by highpass filtering, lowpass filtering, or additive noise. Confusion matrices were determined by recognizing the syllables using different ASR front ends, including Mel-Filter Bank (MFB) energies, Mel-Filtered Cepstral Coefficients (MFCCs), and the Ensemble Interval Histogram (EIH).

For syllables degraded by lowpass and highpass filtering, automated systems trained on the degraded condition recognized the consonants roughly as well as humans. Moreover, all the ASR systems produce similar patterns of recognition errors for a given filtering condition. These patterns differ significantly from that characteristic of humans under the same filtering conditions.

For syllables degraded by additive speech-shaped noise, none of the automated systems recognized consonants as well as humans. As with filtered conditions, confusion matrices revealed similar error patterns for all the ASR systems. While the error patterns of humans and machines was more similar for noise conditions than for filtered conditions, the similarities were not as great as between the ASR systems. The greatest difference between human and machine performances was in determining the correct voiced/unvoiced classification of consonants.

Given these results, work was focused on recognition of the correct voicing classification in additive noise (0 dB SNR). The approach taken attempted to automatically extract attributes of the speech signal, termed sub phonetic features, which are useful in determining the distinctive feature voicing.

Two sub phonetic features, intervocal period (the length of time between the onset of the vowel and any preceding vocalization) and delta fundamental (the average first difference of fundamental frequency over the first 90 msec of the vowel) proved particu-
larly useful. When these two features were appended to traditional ASR parameters, the deficit exhibited by automated systems was reduced substantially, though not eliminated.
Chapter 1

Introduction

Despite significant advances in Automated Speech Recognition (ASR) systems, performance at human levels has not yet been attained. Lippmann [Lippmann 1997] compared results of human and machine speech recognition and found that machine word error rates were typically about an order of magnitude greater in quiet environments. The gap in performance between humans and machines tends to get larger for speech in noise or when the recognition task gets more complex, for example as vocabulary size increases or as the task moves from recognition of isolated words to continuous speech. In this thesis, human and machine performance is compared at the phone recognition level. This extends comparisons to a less complex recognition task than even isolated word recognition.

Human recognition results provide proof that continuous speech can be recognized more reliably than the best current ASR systems. In theory, if a complete model of human speech processing were available, human-level performance would be immediately realizable. While no such model is currently available, knowledge of how humans go about the speech recognition task is potentially useful in directing research on improving ASR systems. The second part of this thesis attempts to use knowledge of human speech processing to improve machine performance at the phone recognition level.
Figure 1-1: Overview of the Automated Speech Recognition (ASR) process. Along the left margin are shown types of knowledge about human audition which can be integrated into ASR systems.

1.1 Automated Speech Recognition

Figure 1-1 shows the stages generally involved in ASR. A sampled speech waveform is first parameterized by the front end of the system. Two examples of traditional parameterizations include Mel-Filter Bank (MFB) energies and Mel-Frequency Cepstral Coefficients.

The back end of the automated system performs some form of pattern recognition on the parameterized speech to produce a sequence of phone labels. Most state-of-the-art systems use Hidden Markov Models (HMMs) for this pattern recognition stage. The sequence of phone labels is often processed by a “higher-level” stage, involving lexical, grammatical, and other constraints to produce the final output of the system, usually words or sentences.
Because recognition results are generally reported as word error rates\textsuperscript{1}, it is unclear how much of the gap in human and machine word recognition performance is due to differences in recognition of basic sounds (phones) and how much is due to differences in how strings of recognized sounds are converted to words (as well as phrases and sentences). This thesis attempts to disambiguate between the performance differences at the phone recognition level and at higher levels. Specifically, human and machine performance on recognition of consonants in nonsense syllables is compared.

Knowledge of performance differences at the phone recognition level can yield insight into which elements of large vocabulary continuous speech recognition systems need to be most improved in order to close the gap between automated and human abilities. For example, if human and machine recognition of phones were shown to be comparable, one could conclude that the gap in word recognition performance arises from superior higher-level processing by humans. Alternatively, it could be shown that the gap in recognition abilities extends to the level of phone recognition, indicating room for improvement of automatic systems at that level.

Performance measurements need not be limited to phone recognition accuracy. Recording confusion matrices for humans and machines on the phone recognition task allows comparison of error patterns. Knowledge of what types of errors are made by machines more frequently than by humans can provide insight into the weaknesses of automated systems. These weaknesses can be used as focal points for work on improving ASR systems. Knowledge of human processing strategies can be used as a basis for improved automated processing.

1.2 Using Knowledge of Human Audition for Automated Speech Recognition

Some ASR researchers have advocated utilizing knowledge of how humans recognize speech in order to improve the performance of automated systems (c.f. [Zue 1985]).

\textsuperscript{1}Ghitza (1993) provides an exception to this which will be discussed later in this chapter.
Along the left side of Figure 1-1 are listed some forms of knowledge about human speech processing that can be incorporated into automated systems. This section describes how knowledge of human speech processing can be integrated into machine systems at the front end and back end levels. Because this research focuses on phone recognition, higher-level processing is not discussed.

1.2.1 Auditory-Based Front Ends

Knowledge of speech encoding at the level of the auditory periphery can be used in selecting or developing waveform parameterizations. The motivation for this approach to parameterization development is that any information that is not represented in the responses of auditory neurons is not necessary for human-level performance. Knowledge of what information is encoded (and possibly emphasized or enhanced) can guide efforts to improve the ability of parameterizations to represent those speech attributes required for human-level recognition performance.

Much work in this area has focused on using models of auditory nerve encoding of speech as the front end or parameterization stage of ASR. The synchrony/mean rate model described in [Seneff 1988] is a prominent example. Seneff developed a two-channel auditory-based model of neural encoding using a rate code in one channel and a temporal (synchrony) code in the other channel. The rate code derives firing probabilities for a number of neural channels, with each channel meant to represent information coded by a number of similarly-tuned auditory neurons. The temporal channel examines the current excitation level of a channel and the delayed excitation level with delay equal to the reciprocal of the frequency being examined. Specifically, the sum of the two excitations (high for firing at the examined delay) is divided by the difference between the two excitations (low for firing at that delay).

Cosi et al. (1990) used a Euclidian distance metric (rather than gaussian probability density estimation) within a Dynamic Time Warping (DTW) algorithm and realized better recognition with cepstral coefficients than with parameters generated by Seneff's model. The use of a DTW algorithm requires knowledge of the start and end points of the basic recognition units (presegmentation) due to endpoint con-
straints (in DTW the input start and end frames and the template start and end frames must be matched respectively, whereas for HMMs the probabilistic nature in effect tries a series of different start and end points). DTW results can not be assumed to reflect results that would be found using the less-constrained HMM approach because it is unclear whether one of the representations would lead to more accurate segmentation by HMM systems.

Hunt & Lefebvre (1988) obtained opposite results when Linear Discriminant Analysis (LDA) was applied to parameters generated by Seneff’s auditory model. Eight LDA coefficients were used to represent a 64-channel combined synchrony and mean rate front end (32 mean rate + 32 synchrony channels) and compared against an 8-element MFCC vector. The speech consisted of three-digit utterances. The MFCC system trained on clean speech displayed 0.22% errors on clean speech, 21% on noisy speech, and 76% on a 6dB/octave emphasis above a few hundred Hz. The auditory-based front end, when trained on clean speech, displayed 0.07% errors on clean speech, 1.3% errors on noisy speech, and 4.5% errors on spectrally-tilted speech.

The poor performance of Hunt & Lefebvre’s cepstral system may have been due to use of an inadequate number of coefficients. Ohshima & Stern (1994) showed that optimal performance using Seneff’s model requires only 4-5 principal components. The use of 8 LDA coefficients could thus be near-optimal for Seneff’s model while 8 MFCCs may have been suboptimal for the MFCC approach (up to 12 MFCCs are commonly used). In a later study, Hunt & Lefebvre (1989) showed that MFCCs gave results comparable to Seneff’s model for clean speech if MFCC coefficients were weighted properly. However, MFCC performance was still not robust to noise or channel mismatch.

Cepstral representations can also be enhanced by the incorporation of $\delta$ parameters (first differences of coefficients from frames typically separated by 40ms). Hunt & Lefebvre (1989) reported that LDA coefficients based on a mel-scale filter bank supplemented with $\delta$ terms were more robust than either an auditory front end based on Seneff’s model or the MFCC parameterization (where neither the auditory front end nor the MFCC was analyzed using LDA). The $\delta$ supplemented mel-filter bank
LDA coefficients still outperformed the MFCC parameterization when MFCCs were supplemented with $\delta$ coefficients. The favorable comparison of the $\delta$-supplemented mel-scale filter bank with Seneff's temporal model (which was not supplemented with $\delta$ values) is interesting in that it indicates that bandpass energy and energy differences are as effective as the much more complicated analysis performed by Seneff's model.

Another auditory-based front end is the Ensemble Interval Histogram (EIH) as described in [Ghitza 1994]. The EIH uses a set of bandpass filters developed to display tuning characteristics similar to neurons in the auditory nerve. After speech is bandpass filtered, threshold detectors produce a 'spike' every time a positive crossing of a threshold occurs. Four (or five) thresholds cover the 30 dB range of normal speech. After converting the speech into a set of spike times on the various channels, the interspike intervals are measured. A histogram of frequencies corresponding to the reciprocals of the intervals is then collected across thresholds and across channels. This histogram is termed the Ensemble Interval Histogram (EIH).

Ghitza (1993) compared performance of cepstral coefficients calculated using the Fourier Power Spectrum with cepstral coefficients calculated using the auditory-based Ensemble Interval Histogram (EIH) front end and with human performance. The speech task was the Diagnostic Rhyme Test (DRT). The DRT is structured to measure relative performance on recognition of phonetic features, which differentiate between consonant sounds. This allowed for comparisons of patterns of responses between humans and automated systems. Recognition was tested for speech in additive white noise at Signal-to-Noise Ratios (SNRs) of +30, +20, and +10 dB.

Human results revealed an overall error rate of roughly 5% at the poorest SNR (+10 dB). In this condition, humans made no errors for the phonetic features voicing and nasality. The systems using the Fourier power spectrum and the EIH had overall error rates of roughly 50% and 25% respectively. While the EIH outperformed the traditional front end in noisy conditions, neither automated system approached human-level performance. The disparity in overall recognition levels and the lack of errors by human subjects for half of the error categories examined makes response pattern comparisons problematic.
Jankowski et al. (1994) compared performance of a mel-cepstra system with Seneff's synchrony/mean-rate model\(^2\) and with Ghitza's EIH in additive noise over a range of Signal to Noise Ratios (SNRs). The speech task was isolated word recognition for a set of 105 aircraft commands. Results from the set of systems are generally similar except at the poorest SNR (+6 dB), where the error rate for the auditory-based systems was roughly 20% compared to 25% for the traditional system.

Human performance levels for recognition of quiet read speech (the CSR'94 Spoke 10 corpus, made up of multiple speakers reading passages from the 5000 word vocabulary Wall Street Journal corpus) do not vary as noise is added down to +10 dB SNR [Ebel & Picone 1995]. Jankowski et al. found that error rates of the automated systems they tested increase from roughly 1.0% in quiet to 7.5% around +12 dB SNR with training on clean speech alone. When training included speech degraded by noise, error rates increased from less than 1% in quiet to roughly 3.5% at +12 dB SNR. This is an example of how machines display less robustness of performance levels in degraded conditions, as well as the sensitivity of machine performance to the training data.

1.2.2 Auditory-Based Back Ends

At the back end, or pattern recognition level, models of central auditory processing could be implemented if they were available. However, detailed physiological studies of processing in the auditory portions of the central nervous system are not advanced enough at the current time to develop usable models for implementation in ASR systems. Alternatively, results of psychophysical studies on human speech recognition that indicate the acoustic cues humans use in making phonetic distinctions can be used to direct development of systems which will automatically extract those attributes.

\(^2\)The tests of Seneff's model separated the synchrony and the mean-rate channels in Seneff's system. It is possible but not proven that a system using both of the channels could show improvement over either of the individual channels.
1.3 Overview of the Thesis

This thesis compares the performance of the auditory-based EIH front end and two traditional front ends, Mel-Filter Bank (MFB) energies and Mel-Frequency Cepstral Coefficients (MFCCs) with data on human performance on the same recognition task. This extends comparisons between humans and machines at the subword recognition level. The task is recognition of consonants in nonsense syllables. The databases used are described in Chapter 2. Following this, the implementation of the ASR systems is described in Chapter 3. Comparisons consider both recognition scores and error patterns to determine which front ends, if any, produce error patterns similar to those of humans. The comparisons are described in Chapter 4.

Based on comparisons of human and automated performance on the CVC recognition task, the determination of consonant voicing in additive noise was identified as a major weakness of ASRs. Previous research on the acoustic differences associated with the voicing distinction and studies of human perception of these differences, identified a set of physical characteristics of the speech signal likely to be salient. These are described in Chapter 5. Chapter 6 describes the method used to automatically extract cues from noisy speech and measures of how effective the individual cues are for voicing classification in 0 dB additive noise. The utility of multiple cues for voicing classification is described in Chapter 7, along with results when the automatically extracted cues were integrated with the traditional ASR systems. Finally, Chapter 8 contains a discussion of the work done in the thesis.
Chapter 2

The Nonsense Syllable Databases

Two sets of human consonant recognition results were used for comparing human and machine recognition. One set of results is for human recognition of a database composed of Consonant-Vowel-Consonant (CVC) syllables. The other set of results is for human recognition of Consonant-Vowel (CV) syllables.

The CVC database is composed of 872 Consonant-Vowel-Consonant syllables, each preceded by a schwa. The syllables were constructed using 12 consonants and 6 vowels. The initial and final consonants for a CVC token were independently drawn from the set of 12 consonants. Table 2.1 lists consonants that are present in the database, along with their classification along four different distinctive feature dimensions.

Two speakers, one male and one female, each produced half of the CVC tokens. Their speech was lowpass filtered at 9 kHz and converted to 12 bit samples at a sampling rate of 20 kHz. The CVC database was later resampled at a 10 kHz sampling rate. This was done in order to match recording characteristics of the TIMIT database ([DARPA 1991]), which was being used for training phone models before testing on the CVC database. In later stages models were trained only on the CVC database but the 10 kHz-sampled tokens were retained.

The CVC database had been used in psychophysical tests of human phone recognition for speech degraded by highpass filtering [Rubin et al. 1998], and by additive speech-shaped noise [Dix 1998]. In order to allow comparison of human and machine recognition of lowpass filtered speech, comparisons were also made with results
Table 2.1: Feature descriptions of the twelve consonants used in the CVC database. For Voicing, Frication, and Sibilence, a 0 indicates the feature is absent and a 1 indicates the feature is present. The place classes follow those of Miller & Nicely: 0 = Labial, 1 = Alveolar, and 2 = Velar. Note that the Sibilence feature here corresponds to Miller & Nicely’s ‘Duration’ feature.

<table>
<thead>
<tr>
<th>Consonant</th>
<th>Voicing</th>
<th>Frication</th>
<th>Sibilence</th>
<th>Place</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>peep</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>toot</td>
</tr>
<tr>
<td>K</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>kick</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>bob</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>deed</td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>gig</td>
</tr>
<tr>
<td>TH</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>thin</td>
</tr>
<tr>
<td>S</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>sis</td>
</tr>
<tr>
<td>SH</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>shin</td>
</tr>
<tr>
<td>V</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>vet</td>
</tr>
<tr>
<td>DH</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>them</td>
</tr>
<tr>
<td>Z</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>zip</td>
</tr>
</tbody>
</table>

reported by Miller & Nicely (1955) using CV syllables.

The Miller & Nicely study used 16 consonants, including /f/, /zh/, /m/, and /n/ in addition to all 12 present in the CVC database. Tests explored recognition in additive white noise, in highpass filtering conditions, and in lowpass filtering conditions. For the highpass and lowpass filtering conditions, speech degraded by additive noise at +12 dB SNR was used, in contrast to the CVC tests which used clean speech for filtering tests.

Because Miller & Nicely only examined consonant confusions (they used only one vowel) and because only 6 vowels are used in the CVC database, only consonant recognition results were considered in this thesis.

During human testing, subjects knew in all instances the constraints of the database (e.g. that whatever token was chosen from the CVC database, it would begin with a schwa and be followed by a Consonant-Vowel-Consonant sequence). Analogously, all of the automated systems were constrained to recognize a period of initial silence (recordings did not start simultaneously with the schwa) followed by a Vowel-
Consonant-Vowel-Consonant sequence of phones (the initial Vowel corresponds to the schwa preceding the CVC sequence), and ending with another interval of silence.

For tests of speech degraded by additive speech-shaped noise, a Grason-Stadler noise generator (model 901B) was used to generate noise with a spectrum that was flat to 1 KHz and then fell at around 6dB/octave. Noise was added to the CVC tokens to create databases at SNRs of 30 dB, 20 dB, 10 dB, 5 dB, and 0 dB\(^1\). For filtering tests, digital filters were designed using Matlab [Mathworks Inc.] software. All filters used 512 coefficients to create sharp cutoffs at the appropriate frequencies (equivalent to roughly a sixth-order highpass or lowpass filter) \(^2\). After filtering, noise (at + 30 dB SNR) was added to the filtered signal to mask the signal in the stop band.

For training purposes, the CVC database was split into three token sets, labeled DB1, DB2, and DB3. The token sets were selected such that both initial and final consonants were evenly distributed across the three sets. ASR systems were trained on two of the three sets and tested on the third (e.g. trained on DB1 and DB2, and tested on DB3). For each front end and for each training approach, three separate instantiations of the ASR system (the particular front end matched with the HMM back end) were trained and tested so that each token set was used once as the test set. The results from the three separate test sets were then averaged to generate the results presented in Chapter 4.

---

\(^1\)To generate a signal at a desired SNR, the Root-Mean-Squared (RMS) energies of all CVC tokens were calculated over entire CVC recordings and then noise was added at an appropriate RMS level relative to the overall average RMS energy. This contrasts with determination of SNR by adding noise relative to peak signal amplitudes or relative to vowel RMS energies. Additionally, the average RMS level of all 872 CVC tokens was used to determine the level of noise added to all of the tokens to achieve a particular SNR (this is consistent with the method used for the human CVC studies). Miller & Nicely did not use recorded CV syllables. Instead, speakers were trained to produce CV syllables at a consistent amplitude. The average peak amplitude over a number of examples was then used as the basis for calculating necessary noise levels for the range of SNRs tested.

\(^2\)Specifically, the FIR2 function was used. This function designs filters using the window method. It takes as input a vector of frequencies and target amplitude for those frequencies. For a particular cutoff frequency (Fc), filters used a frequency vector of \([0 \ 0.95 Fc \ Fc \ 1.05 Fc \ 1]\). Corresponding amplitude vectors were \([0 \ 0.7 \ 1 \ 1]\) for highpass filters and \([1 \ 1.7 \ 0 \ 0]\) for lowpass filters.
Chapter 3

The Automated Recognition Systems

Three front ends: Mel-Filter Bank (MFB) energies, Mel-Filter Cepstral Coefficients (MFCCs), and the Ensemble Interval Histogram (EIH), were paired with a Hidden Markov Model (HMM - [Rabiner & Juang 1993]) back end. Following discussion of the ASR systems, the three different training approaches which were tested on the systems will be described.

3.1 The MFB and MFCC Parameterizations

The MFB and MFCC front ends were implemented using the HTK software package (Entropic Research Laboratory, Inc.). Each of these front ends generated a set of parameters every ten milliseconds using a twenty millisecond Hamming window for analysis of the input signal. The ten millisecond frame rate and twenty millisecond analysis window values are within the ranges typically used for ASR systems (c.f. [Jankowski et al. 1994]).

The Mel-Filter Bank (MFB) front end derived spectral energies using a mel-warped frequency axis:

\[ Mel(f) = 2595 \log_{10}(1 + f/700) \]
and a set of 24 triangular filters equally spaced to cover the mel-warped axis. Each filter had unit magnitude at its center and extended to the center (peak) of the neighboring filters. The 24 spectral energies were supplemented by 24 δ-coefficients, weighted measures of the rate of change for each spectral energy parameter over a 50 msec time span. In addition, an energy term was calculated. This energy value was only used to calculate a delta value (corresponding to change in overall energy) and was not used directly as a parameter. This yielded a final 49-element parameter vector (24 MFB values, 24 δ-MFB values, and 1 δ-energy value).

The MFCC front end used a Discrete Fourier Transform to convert the mel-warped spectra into 13 cepstral coefficients for each frame of speech. The zeroth cepstral coefficient (corresponding to the signal energy in the frame) was used in the same way as the energy value in the MFB system. This resulted in a final 25-element parameter vector (12 cepstral coefficients, 12 δ-cepstra, and 1 δ-energy value).

Preliminary testing explored the performance of the MFB and MFCC parameterization paired with HMMs using different numbers of states and mixtures per state. The numbers of states and mixtures for the HMM showing the highest accuracy on clean speech were used for all further tests of the parameterizations. For the MFCC front end, each phone was modeled with a three-state left to right (ergodic) HMM. Within each state, a mixture of 12 diagonal-covariance gaussians was used to model output probabilities. For the MFB front end, each phone model had 5 states and each state used 15 gaussians to model output probabilities.

### 3.2 Ensemble Interval Histogram (EIH)

The EIH front end [Ghitza 1994] can be divided into the stages of waveform filtering, determination of individual ‘firings’ in model channels, and calculation of the intervals between model firings. This produces a spectral representation of the speech signal based upon inter-‘spike’ intervals in the model neural channels.

The filters used in the implementation were generated using a Matlab toolkit [Slaney 1994]. Figure 3-1 shows a subset of the 73 filters which were used.
Figure 3-1: Frequency response for 24 (of 73) filters generated with the Matlab auditory toolkit based on Lyon’s passive ear model.

The output of the filtering stage resulted in a 73-channel representation of the speech input. Each channel was then half-wave rectified. The next stage determined the times that a positive threshold crossing occurred for each of seven logarithmically-spaced thresholds for each channel. The original EIH representation varied the thresholds for each channel of the filtering output to match amplitude ranges in individual channels. In this research a single set of thresholds was used for all channels in order to reduce computational complexity. The thresholds were logarithmically spaced to cover the range of amplitudes exhibited by low-frequency channels for some sample CVCs. This resulted in a 511 bit representation (73 channels X 7 thresholds per channel) where each sample had a value of 1 (when the threshold for that channel had been crossed positively) or zero (when the threshold had not been crossed positively).

Intervals between consecutive firings in individual channels were then recorded. To determine the EIH representation for a particular frame, each channel was examined over a length of time equal to 10 times the reciprocal of the channel's Center Frequency (CF, calculated as the frequency corresponding to the peak in the frequency response curve for the filter used to generate the channel) back in time with a maximum window length of 40 msec.

A histogram with 24 bins (chosen to be equivalent to the number of filters used in the MFB representation) was constructed based on the number of samples in the intervals between firings. The bins were determined by logarithmically spacing fre-
quency thresholds between roughly 100 and 5000 Hz with some constraints added to the procedure. In order not to increase the computational load by requiring up-sampling or interpolation, it was decided that bins would be rounded to an integer number of samples (the waveforms were sampled at 10 KHz, so samples were spaced 0.1 msec apart). An obvious additional constraint was that no two bins round to the same number. Table 3.2 shows the sample intervals per bin. These constraints were not used by Ghita in his original EIH implementation. Their effect is to give more weight to the low frequencies than either Ghita's original EIH or either of the two traditional parameterizations (MFB or MFCC).

Table 3.1: Inter-'spike' intervals for each EIH bin.

<table>
<thead>
<tr>
<th>Bin #</th>
<th>Samples</th>
<th>Bin #</th>
<th>Samples</th>
<th>Bin #</th>
<th>Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>85-97</td>
<td>9</td>
<td>26-29</td>
<td>17</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>73-84</td>
<td>10</td>
<td>23-25</td>
<td>18</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>63-72</td>
<td>11</td>
<td>20-22</td>
<td>19</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>55-62</td>
<td>12</td>
<td>17-19</td>
<td>20</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>47-54</td>
<td>13</td>
<td>15-16</td>
<td>21</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>41-46</td>
<td>14</td>
<td>13-14</td>
<td>22</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>35-40</td>
<td>15</td>
<td>11-12</td>
<td>23</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>30-34</td>
<td>16</td>
<td>10</td>
<td>24</td>
<td>2</td>
</tr>
</tbody>
</table>

The EIH, with δ-parameters calculated identically to the MFB parameterization\(^1\), was matched with a Hidden Markov Model back end. A preliminary set of tests explored recognition performance on the clean CVC database using different numbers of HMM states and gaussian mixtures. It was decided that a 3-state 12-mixture system would be used because it was the smallest system which displayed asymptotic-level performance (a number of systems with larger numbers of states and/or mixtures displayed equivalent performance).

\(^1\)Note that the δ-parameters were consistently calculated using samples spaced 10 msec apart, even though the analysis windows for the various channels ranged from around 2-40 msec. It is possible that EIH performance could be improved through optimization of the algorithm for calculating δ-parameters. However, as the main thrust of this research was the exploration of response patterns, it was not deemed worthwhile to optimize every element of the EIH implementation.
3.3 Training Approaches

The performance of ASR systems, especially in degraded signal conditions, is heavily dependent on the speech used to train the systems. Systems trained and tested under unmatched conditions have displayed poor performance. Three types of training and testing regimens were used to evaluate system robustness:

- Clean Training (CT): A system trained only on clean, broadband speech is tested in a variety of degraded conditions.

- Multi-Style Training (MST): A system trained over a range of degradations (e.g. speech degraded by additive noise at 0, 5, 10, 20, and 30 dB SNR) is tested over the same range (0, 5, 10, 20, and 30 dB SNR).

- Degradation-Specific Training (DST): A system is trained and tested on the same degradation (e.g. additive noise at 5 dB SNR).

To examine how well a CT system performs on the various degradations, it was only necessary to train a single system on the undegraded database and then test it across the entire set of degradations (both filtering and noise). For MST testing, a separate system is trained for each type of degradation (e.g. highpass filtering) and tested separately for each degree of severity for that degradation (e.g. 200, 400, 800, 1600, and 2800 Hz highpass cutoff). For DST testing, a separate system needs to be trained for each type and severity of degradation (e.g. highpass filtering with a 200 Hz cutoff frequency)².

Human training is hard to match to any of the above training regimens. Because most speech is heard in a variety of noisy environments, the Clean Training method seems unreasonable. The Multi-Style training might be most applicable, however the

²This description is somewhat simplified in that it ignores the fact that separate systems were trained for testing on each of the three subdivisions of the database. There were in fact three systems trained according to the CT approach, each using a different subdivision of the CVC database as its test set and the rest of the CVC database as its training set, and the results reflect the average of the three separate training and testing sets for a particular training approach. For systems trained in multiple degradations (e.g. trained on all noise levels) the same set of tokens was used across the range of degradations (e.g. the token 'bahb' would only be in the training set or the testing set, not the training set at some degradations and the testing set at other degradations).
set of environments a typical human is likely to encounter is unlikely to include all of
the degraded conditions that have been tested. Nevertheless, humans perform very
well even in these “untrained” conditions (e.g. highpass filtering at 2 kHz). This
suggests an ability to handle or adapt to novel environments. It is also possible that
humans recognize speech using a number of different recognition criteria dependent
upon their perceived environment. If this is the case, then the Degradation-Specific
Training is likely the best model. The actual training approach used by humans is
probably somewhere between the extremes listed above.

CT evaluates the baseline performance of the different systems and determines how
sensitive they are to different degradations. Performance of CT systems on degraded
speech is expected to be worse than the performance of MST systems, which will
have been trained on samples of the degraded speech (c.f. [Jankowski et al. 1994]).
At a particular degradation, (e.g. additive noise at +10dB SNR), a DST system is
expected to outperform either a multistyle-trained system or a clean-trained system.
This is because the DST approach allows the system to utilize all of its parameters
to model speech at that degradation, where the same number of parameters will be
used to model speech at a range of degradations in the MST approach or a different
set of speech examples for the CT approach.

The three approaches to training were utilized for both additive noise and filtering
degradations of the speech signal, though degradations were never combined (e.g.
there were no tests of speech degraded by additive noise and highpass filtering).
Chapter 4

Comparisons of Human and Machine Consonant Recognition

In this chapter, human performance on the CVC and CV databases is compared against machine performance on the CVC database for highpass filtering, lowpass filtering, and additive noise degradations. Phone error rates are first compared. Next, Information Transfer scores are reported for divisions of the phone set based on phonetic features. Finally, confusion matrices are examined using a Metric Multidimensional Scaling (MMS) method to compare phone pair confusability displayed by humans and machines.

4.1 Error Rate Comparisons

4.1.1 Highpass Filtering Results

Figure 4-1 compares human results on highpass filtered speech with results from the MFB automated system. Because Miller & Nicely used phones degraded by +12 dB SNR additive noise for their filtering studies, their results are expected to be somewhat lower than results using filtered versions of clean speech. Three sets of MFB results are plotted, one for each of the three training methods used.

CT accuracy is consistently the lowest over the range of highpass conditions. The
Figure 4-1: Percent correct scores for humans and the MFB systems on speech degraded by highpass filtering.

MST results are much better than CT results. MST accuracy approaches human-level performance. Finally, DST results are the best, slightly superior to the MST results and again showing human levels of performance.

Figures 4-2 and 4-3 show highpass filtering results for the MFCC and EIH systems respectively. Consistent with the MFB results, Clean Training yields the worst performance and Degradation-Specific Training yields the best performance.

Figure 4-4 displays human scores with the DST results from the three automated systems. The MFB and MFCC data points are very similar. Both display human levels of performance, even surpassing human performance at the most severe highpass condition (2800 Hz cutoff). The EIH system does not do as well in highpass conditions. This is likely due at least in part to the constraints placed on interspike interval bins, which provided fewer separate bins for high frequency channels when compared to the MFB and MFCC systems. EIH results never reach human levels although the gap in performance between humans and the EIH system decreases as the highpass cutoff is increased from 1000 Hz to 2800 Hz.
Figure 4-2: Percent correct scores for humans and the MFCC systems on speech degraded by highpass filtering.

Figure 4-3: Percent correct scores for humans and the EIH systems on speech degraded by highpass filtering.
4.1.2 Lowpass Filtering Results

Figure 4-5 compares results of human and MFB tests for lowpass filtered speech. In addition to Miller & Nicely's CV results, predicted results on the CVC database are presented. These predictions are made using an Articulation Index calculation (the 20-band method, [ANSI 1970]). Predictions are based on the presence of equal numbers of contributing frequency bands as were present for the CVC highpass results. Once again it should be noted that Miller & Nicely used CV tokens degraded by +12 dB SNR additive noise.

As was found in the highpass tests, CT accuracy is consistently the lowest. The MST and DST recognition accuracy shows roughly human-level performance. The peak in MST accuracy between 2 and 3 KHz cutoff frequencies is likely due to the fact that four different lowpass cutoffs were all contained in that range (data points at lowpass cutoffs of 2.0, 2.4, 2.7, and 3.0 KHz) so that over half of the training data came from this range of cutoffs.

Figures 4-6 and 4-7 display lowpass results for the MFCC and EIH systems re-
Figure 4-5: Percent correct from human CV results, predicted CVC scores, and MFB systems for speech degraded by lowpass filtering.

drespectively. All three machine systems show comparable results for lowpass conditions, with roughly human-level performance when training data included the lowpass condition being tested.

Figure 4-8 shows results for humans and the three automated systems using Degradation-Specific Training. The MFB and MFCC results largely overlap at cutoff frequencies above 2000 Hz, with the MFB results slightly superior to the MFCC results at lower cutoffs. The EH front end consistently results in accuracy 5-10 percentage points below MFB and MFCC.

4.1.3 Additive Noise

Figure 4-9 compares results of human and MFB tests for speech degraded by additive noise. Human results are shown for recognition in additive white noise (Miller & Nicely's CV results) and for recognition in additive speech-shaped noise (Dix's CVC results). In addition to recognition results, predicted accuracy based on Articulation Index theory is included. The predicted scores largely overlap the human CVC results,
Figure 4-6: Percent correct from human CV results and MFCC systems for speech degraded by lowpass filtering.

Figure 4-7: Percent correct from human CV results and EIH systems for speech degraded by lowpass filtering.
Figure 4-8: Percent correct scores for humans and the automated systems using Degradation-Specific Training (DST) on speech degraded by lowpass filtering, validating the use of predicted CVC results for comparison with machine results in the lowpass filtering condition.

The two sets of human recognition results for additive noise largely overlap despite the different spectral characteristics of the additive noise used (white noise for CV tokens, speech-shaped noise for CVC tokens). Speech-shaped noise was used for all ASR tests.

Whereas for both of the filtering conditions the MST and DST approaches yielded human-level recognition rates, for speech degraded by additive noise machine recognition accuracy is 15 to 20 percentage points lower than human accuracy. Figures 4-10 and 4-11 show similar results for the MFCC and EIH systems. CT results are consistently the poorest in additive noise. Generally, the DST results are the best though in a few cases MST yields greater accuracy than DST.

Figure 4-12 compares DST results for the three machine systems with human performance on speech degraded by additive noise. The EIH front end consistently yields the poorest results. The performance difference between the EIH front end
Figure 4-9: Percent correct scores from human data and MFB systems for speech degraded by additive noise.

Figure 4-10: Percent correct scores from human data and MFCC systems for speech degraded by additive noise.
Figure 4-11: Percent correct scores from human data and EIH systems for speech degraded by additive noise.

and the other front ends is not as large as for highpass-filtered speech, being more similar to the results for lowpass-filtered speech. The MFB system is consistently better than the MFCC system except at 30 dB SNR, where the two points overlap. In contrast with the results for highpass and lowpass filtering, none of the machine systems displays human-level performance for speech degraded by additive noise. The machine systems require 5 to 10 dB higher SNRs to reach human accuracy.

4.2 Information Transfer Analysis of Confusion Matrices

Information Transfer (IT) scores were calculated for four divisions of the phone set corresponding to the phonetic features of voicing, frication, place, and sibilance as shown in Table 2.1. As in the analysis of error rates, the three automated systems displayed similar patterns of IT scores. Only MFB IT scores will be compared with
Figure 4-12: Percent correct scores for humans and the automated systems using Degradation-Specific Training (DST) on speech degraded by lowpass filtering.

human results because of the similarity of scores across automated systems. Plots of IT scores for MFCC and EIH systems can be found in Appendix A.

4.2.1 Highpass Filtering Results

Figure 4-13 shows four sets of IT scores based on four different phonetic features. The first plot compares sibilance IT scores for humans on the CVC and CV databases with MFB results for each of the three training methods. Both MST and DST results are well above human levels when the highpass cutoff is above 1000 Hz. A similar situation is evident for frication IT scores. While place IT scores for the MST and DST systems are consistently above human scores, the differences between machine and human scores are not as large as for sibilance and frication. Voicing IT scores for the machine systems are similar to human scores.

Human IT scores are fairly consistent across the four features examined. In contrast, machine systems show a greater ability to correctly determine the sibilance and frication classes compared to place and voicing classes.
Figure 4-13: Information Transfer (IT) scores from human CVC and CV data and from MFB systems on highpass filtered speech.
4.2.2 Lowpass Filtering Results

Figure 4-14 shows IT scores for lowpass-filtered speech. Sibilance scores for the MST and DST machine systems are again well above human levels. MST and DST frication and place scores show a similar strength for the machine systems, though the differences between machine and human scores are not as great as for sibilance. For voicing IT scores, however, humans are performing at the level of the best machine systems despite the gap in phone recognition scores, showing a greater robustness than for the other three features.

In contrast to highpass results, humans show varying abilities for recognizing each of the four features. Voicing is the most consistently recognized feature especially as the cutoff frequency is lowered. In contrast to the human trend, the automated systems show similar IT score patterns in lowpass and highpass conditions.

4.2.3 Additive Noise Results

The final set of IT score comparisons is for speech degraded with additive noise. In Figure 4-15 are displayed human results using the CV database with additive white noise [Miller & Nicely 1955] and the CVC database with additive speech-shaped noise [Dix 1998]. IT scores for the Place feature largely overlap for the two human experiments, and frications scores are similar. The differences between the two sets of human results are in the voicing and sibilance features. Tests using white noise resulted in higher voicing IT scores and lower sibilance IT scores when compared with tests using speech-shaped noise. This is likely due to the spectra of the respective noise maskers, with white noise having relatively more high-frequency energy than the speech-shaped noise. Since sibilance information largely resides in the high-frequency portion of the speech spectrum, at a given SNR white noise is likely to be more deleterious than speech shaped noise. If voicing is based on the low-frequency portion of the speech spectrum, the white noise would be less deleterious than speech-shaped noise at the same SNR.

Sibilance IT scores are comparable for human and machine systems using speech-
Figure 4-14: Information Transfer (IT) scores from human CV data and from MFB systems on lowpass-filtered speech.
MFB/Human Noise Comparison: IT Scores

Figure 4-15: Information Transfer (IT) scores from human CVC and CV data and from MFB systems on speech degraded by additive noise.
shaped noise despite humans’ overall superiority in noise. Comparisons of the frication
and place IT scores are consistent with the roughly 5 to 10 dB shift needed to match
percent correct scores of humans and the MFB system (see Figure 4-9). Voicing IT
scores suggest the need for a roughly 10 to 15 dB shift required to match MFB scores
with humans. This reveals voicing classification as a major weakness of the machine
systems compared to humans.

4.3 Comparison of Consonant Pair Confusions

4.3.1 Metric Multidimensional Scaling (MMS)

The next form of analysis compares the level of confusability for consonant pairs using
a type of Metric Multidimensional Scaling [Braida 1991]. The Metric Multidimen-
sional Scaling comparison is finer grain than the comparison based on Information
Transfer scores. Information Transfer scores are dependent upon the features revealed
by different divisions of the phone set. In MMS, all possible confusions are compared
across the two conditions.

The MMS method proposes that a stimulus presentation generates a D-dimensional
vector \( \vec{S} = \langle s_1, s_2, s_3, ..., s_D \rangle \) in the perceptual space of the listener. Each stim-
ulus class (in the analysis performed here, each of the twelve consonantal phones
is a separate stimulus class) has associated with it a point in the multidimensional
perceptual space around which the perceptual vectors for instances of that stimulus
are normally distributed. The coordinate system used to model the perceptual space
is scaled relative to the common standard deviation of the gaussian densities. The
differentiability of two stimuli is determined by the distance between the two stimu-
lus centers \( d' \) within the modeled perceptual space. Larger \( d' \) values correspond to
better differentiability (and therefore fewer confusions). \( d' \) values around 5 and above
all correspond to nearly perfect differentiability (one or no confusions) for the number
of stimulus presentations examined in this study and may be treated as equivalent.

In order to account for biases in response patterns, each stimulus also has asso-
associated with it a response center. When a stimulus is presented, the resulting vector in perceptual space is compared against the set of response centers. The response center that is closest to a stimulus vector determines the response that the stimulus elicits. This effectively divides the D-dimensional perceptual space into a number of compact regions equal to the number of response classes. The probability of a correct response given a stimulus from class \( i \) is then the integral of the probability of the stimulus vector for class \( i \) resulting in a vector that falls in response region \( i \).

An iterative procedure was used to estimate a set of stimulus and response centers that would fit a given confusion matrix for a particular dimensionality of perceptual space. All of the results presented used a 4-dimensional perceptual space to balance prediction accuracy with calculation costs. The perceptual distances between stimulus centers (in \( d' \) units) were then determined for each pair of consonants. Note that the specific coordinates of stimulus centers in perceptual space are irrelevant; only the distances between stimulus center pairs are relevant for determining how differentiable the two consonants were. Due to limitations of the iterative procedure, only confusion matrices with a significant number of confusion entries could be used to construct a set of stimulus and response centers. This limited the Metric Multidimensional Scaling analysis to conditions with roughly 80% or fewer correct responses. Comparisons of the \( d' \) measures for pairs of consonants were made between various automated and human systems.

An example comparison is given in figure 4-16. This scatter plot compares consonant pair \( d' \) measures between two different training approaches (MST and DST) for a cepstral system recognizing speech degraded by highpass filtering with a cutoff frequency of 1600 Hz. Each data point represents a particular pair of consonants. The \( x \)-coordinate plots the \( d' \) value for the MST system while the \( y \)-coordinate plots the \( d' \) value for the DST system.

If all of the data points fell along the main diagonal then the \( d' \) measures for all consonant pairs would be equivalent for the two systems (corresponding to equivalent consonant pair differentiabilities for the two recognizers, and therefore roughly equivalent confusion matrices). In fact, if the data points fell along the diagonal for all
Figure 4-16: $d'$ scores for MST and DST cepstral systems recognizing speech highpass filtered with a 1600 Hz cutoff frequency.

points with either coordinate below around 5 $d'$ units, the confusion matrices should be considered equivalent because higher $d'$ values all correspond to consonant pairs that are not confused.

4.3.2 The Scaling Model

One method of modeling the effect of varying the severity of a particular degradation is to uniformly scale the distances between all of the stimulus centers. If this scaling approach is valid, then comparing consonant pair $d'$ values calculated from results of two conditions using the same type of degradation but at different severities (e.g. additive noise at +20 SNR and +10 SNR) should result in data that fall on a line with non-unit slope. The slope of the line then describes the scaling factor. The practical effect of the scaling model is that there is a consistent stimulus center distribution valid for predicting perception of speech for a particular degradation type. Knowing the set of stimulus centers at one severity would then allow prediction of confusion matrices for other severities if the conversion from severity change to scaling factor is
known.

4.3.3 MMS Analysis of Phone Confusion Matrices

For this investigation, we report correlation coefficients for regression analysis between consonant pair $d'$ measures for confusion matrices. The correlation coefficient measures how linear the relationship is between two sets of $d'$ measures. Correlation coefficient scores vary between -1 and 1, with higher absolute values indicating a more linear relationship between two sets of data. The MST vs DST comparison shown in 4-16 had a correlation coefficient of .92, indicating similarity in the response patterns of MST and DST systems.

This analysis provides a method of determining how valid the scaling theory is for sets of confusions matrices describing a degradation type (e.g. a set of additive noise levels). When used to compare data from different machine systems, or from human and machine systems, it describes how similar the relative pairwise distances between stimulus centers are for the systems used to generate the confusion matrices.

Regression analysis requires that the sets of $d'$ values used vary over some range. The restriction of selecting for analysis only conditions which exhibited a large number of non-zero entries in the confusion matrices ensures that data sets will not be limited to large $d'$ values (over 5.0). A further limitation of not using MMS for conditions with low (less than 30%) percent correct scores helps to ensure that there will be a spread of $d'$ values for all conditions, so that during comparison there won't be any sets of $d'$ measures that don't have both low and high values.

For ASR data, the degradations selected for analysis include two highpass filtering conditions with cutoffs of 1600 and 2800 Hz, two lowpass filtering conditions with cutoffs of 500 and 1000 Hz, and three additive noise conditions with 0, +5, and +10 dB SNR. Human CVC results were used for both highpass and additive noise comparisons, while CV results were used for comparisons involving human performance on lowpass filtered speech. For human results, three highpass CVC conditions were used with cutoff frequencies of 1400, 2100, and 2800 Hz. The three CVC additive noise conditions used SNRs of -7, -4, and +2 dB. The CV data used for lowpass results
Table 4.1: Regression analysis correlation coefficients, standard deviations, and numbers of comparisons for consonant pair comparisons of the automated systems using Multi-Style and Degradation-Specific Training.

<table>
<thead>
<tr>
<th>Front End</th>
<th>Mean Correlation</th>
<th>σ</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFB</td>
<td>.93</td>
<td>.04</td>
<td>7</td>
</tr>
<tr>
<td>MFCC</td>
<td>.92</td>
<td>.05</td>
<td>7</td>
</tr>
<tr>
<td>EIH</td>
<td>.91</td>
<td>.06</td>
<td>7</td>
</tr>
<tr>
<td>All Systems</td>
<td>.92</td>
<td>.05</td>
<td>21</td>
</tr>
</tbody>
</table>

included cutoff frequencies of 600, 1200, and 2500 Hz.

Table 4.1 shows average correlation coefficient scores describing the similarity of Multi-Style and Degradation-Specific training approaches. The high average correlation value, .92, and relatively small standard deviation, .05, indicate that the response patterns for automated systems were similar whether Multi-Style or Degradation-Specific training was used. This is consistent with the large amount of overlap in MST and DST error rates and IT scores.

The next comparison examines the different parameterizations using Degradation-Specific Training (Table 4.2). For each pair of parameterizations (MFB vs MFCC, MFB vs EIH, and MFCC vs EIH), a separate regression analysis was performed for each of the seven degradations being examined. Again there is a high average regression value with a relatively small standard deviation, indicating similarity in the response patterns produced by the various automated systems given a particular degradation. This shows that the machine systems' response patterns were similar regardless of which front end was used.

The similarities of response patterns across ASR systems suggests that the systems are using similar aspects of the speech signal for consonant classification. It appears unlikely that a hybrid system (e.g. utilizing both MFB and EIH parameterizations) would outperform the original systems.

Comparisons of the recognition of initial consonants with recognition of final con-
Table 4.2: Regression analysis values for consonant pair comparisons of the different front ends.

<table>
<thead>
<tr>
<th>Consonant</th>
<th>MFB vs MFCC</th>
<th></th>
<th>MFB vs EIH</th>
<th></th>
<th>MFCC vs EIH</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>σ</td>
<td>n</td>
<td>Mean</td>
<td>σ</td>
</tr>
<tr>
<td>Initial</td>
<td>.91</td>
<td>.04</td>
<td>7</td>
<td>.88</td>
<td>.06</td>
</tr>
<tr>
<td>Final</td>
<td>.93</td>
<td>.05</td>
<td>7</td>
<td>.87</td>
<td>.06</td>
</tr>
</tbody>
</table>

Table 4.3: Regression analysis values for consonant pair comparisons: Initial Consonant Recognition vs Final Consonant Recognition.

<table>
<thead>
<tr>
<th>Condition</th>
<th>ASR</th>
<th></th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>σ</td>
<td>n</td>
</tr>
<tr>
<td>Highpass</td>
<td>.79</td>
<td>.06</td>
<td>12</td>
</tr>
<tr>
<td>Lowpass</td>
<td>.86</td>
<td>.06</td>
<td>12</td>
</tr>
<tr>
<td>Noise</td>
<td>.86</td>
<td>.05</td>
<td>18</td>
</tr>
</tbody>
</table>

Table 4.3. For each degradation severity examined (2 for filtering degradations, 3 for the additive noise degradation), a single comparison was made for each of three parameterizations for each of two training approaches (MST and DST). Results show more variability in the response patterns for initial and final consonants than the amount of variability between different front ends. For human results, only highpass and noise comparisons can be made because the lowpass results exist only for the CV database. Human and automated correlations are equal for highpass conditions. For noise conditions, machine systems show more consistency in response patterns of initial and final consonants compared to humans.

The similarity of response patterns across different types of degradation was examined by comparing consonant pair d' scores between systems having roughly 50% correct scores but different degradation types. Results are shown in Table 4.4. For automated systems, d' scores from systems recognizing speech degraded by highpass
Table 4.4: Regression analysis values for comparisons of response patterns on the different signal degradations.

<table>
<thead>
<tr>
<th>Condition</th>
<th>ASR Results</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean Correlation</td>
<td>σ</td>
<td>n</td>
</tr>
<tr>
<td>Highpass vs Lowpass</td>
<td>.47</td>
<td>.10</td>
<td>6</td>
</tr>
<tr>
<td>Highpass vs Noise</td>
<td>.84</td>
<td>.04</td>
<td>6</td>
</tr>
<tr>
<td>Lowpass vs Noise</td>
<td>.58</td>
<td>.13</td>
<td>6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Condition</th>
<th>Human Results</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean Correlation</td>
<td>σ</td>
<td>n</td>
</tr>
<tr>
<td>Highpass vs Noise (CVC)</td>
<td>.16</td>
<td>.15</td>
<td>2</td>
</tr>
<tr>
<td>Highpass vs Noise (CV)</td>
<td>.25</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Highpass vs Lowpass (CV)</td>
<td>.06</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Lowpass vs Noise (CV)</td>
<td>.76</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

filtering with a 2800 Hz cutoff, lowpass filtering with a 500 Hz cutoff, and additive speech-shaped noise at 0 dB SNR were compared. Results from initial consonant recognition and final consonant recognition were compared separately across the three parameterizations for the Degradation-Specific Training approach.

The similarity measure was highest for comparison of additive noise and highpass filtering. This is reasonable since speech-shaped noise tends to mask information contained in the low-frequency region. The other two comparisons (highpass vs lowpass, lowpass vs noise) showed low correlations. Figure 4-17 provides a comparison of results using MFCC DST results on lowpass filtered speech with a cutoff of 500 Hz and noisy speech at a SNR of 0 dB, with a .58 correlation score. In the graph, the additive noise d’ scores show a bimodal distribution while the lowpass d’ scores show a more unimodal distribution. The two distributions for the noise scores correspond to those consonant pairs which contain sibilant sounds (the high d’ scores, corresponding to few errors) and those consonant pairs which do not contain sibilant sounds (the low d’ scores, corresponding to many errors). As shown by the graph, correlation scores at these values (and below) indicate little similarity in relative consonant pair differentiability.

Human cross-degradation comparisons include fewer numbers of confusion matrix
Figure 4-17: d' scores for MFCC systems recognizing speech degraded by lowpass filtering with a 500 Hz cutoff frequency and additive noise at 0 dB SNR.

comparisons. In three of four comparisons, results are based on a single comparison of confusion matrices. For the CVC data, a comparison can only be made between highpass and noise conditions because no lowpass results are available. The comparison between highpass and noise conditions revealed correlation scores of 0.27 and 0.05 for the initial and final consonants respectively. For the CV data, comparisons were made between results from speech degraded by lowpass filtering with a 600 Hz cutoff frequency, highpass filtering with a 2000 Hz cutoff frequency, and additive noise at 0dB SNR. Comparison of highpass results with lowpass and noise results gave correlation values of 0.06 and 0.25 respectively. In contrast, comparison of lowpass and noise results gave a correlation value of 0.76. It seems that for humans, highpass filtered speech results in a fundamentally different response pattern when compared with either lowpass filtered or additive-noise degraded speech (which is consistent with Miller & Nicely's original conclusions for speech degraded by white noise.) For machines, lowpass filtered conditions result in a fundamentally different response pattern compared to highpass filtering or additive speech-shaped noise.
Table 4.5: Regression analysis values for comparisons of response patterns for varying severities of the different types of signal degradation.

<table>
<thead>
<tr>
<th>Condition</th>
<th>ASR Results</th>
<th></th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean Correlation</td>
<td>σ</td>
<td></td>
</tr>
<tr>
<td>Highpass</td>
<td>.85</td>
<td>.04</td>
<td>12</td>
</tr>
<tr>
<td>Lowpass</td>
<td>.90</td>
<td>.04</td>
<td>12</td>
</tr>
<tr>
<td>Noise</td>
<td>.92</td>
<td>.04</td>
<td>36</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Condition</th>
<th>Human Results</th>
<th></th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean Correlation</td>
<td>σ</td>
<td></td>
</tr>
<tr>
<td>Highpass (CVC)</td>
<td>.88</td>
<td>.05</td>
<td>3</td>
</tr>
<tr>
<td>Lowpass (CV)</td>
<td>.32</td>
<td>.50</td>
<td>3</td>
</tr>
<tr>
<td>Noise (CVC)</td>
<td>.93</td>
<td>.01</td>
<td>3</td>
</tr>
</tbody>
</table>

Results of comparisons made across severity for each of the three degradation types are found in Table 4.5. For ASR systems on lowpass and highpass degradations, there was a single comparison (lowpass 500 vs lowpass 1000 and highpass 1600 vs highpass 2800). For noise, there were three comparisons (three combinations of SNR 0, SNR +5, and SNR +10). For each severity, MST and DST results were separately compared. Scores are high (.85 and higher) with relatively small standard deviations (.04 for all three degradation types). These high correlation values come despite up to a roughly 30% difference in percent correct across the worst noise condition (SNR 0) and best noise condition (SNR 10) for the MFCC system. Thus the scaling model seems to hold for changes in severity for the automated systems.

Human data were also examined for cross-severity comparisons. There were three comparisons possible for each of the three degradation types. Results from the human data and the automated data are similar to one another for cross-severity comparisons of highpass-filtered and of noise-degraded speech, supporting the scaling model.

When examining human lowpass (CV) results, low similarity scores resulted when results from the 2500 Hz lowpass cutoff filtering condition were compared with results from lowpass cutoffs of 1200 or 600 Hz. Correlation coefficients for these comparisons were .05 and .01 respectively, indicating very little similarity in confusion matrices.
In contrast, when the 600 and 1200 Hz cutoff data were compared against each other a .90 regression measure resulted. This indicates a fundamental difference in the response patterns of the human subjects when tested on lowpass filtered speech with a 2500 Hz cutoff compared to 600 or 1200 Hz cutoffs, and contradicts the scaling model. A potential explanation of the different response patterns is that the 2500 Hz lowpass condition response pattern is dominated by the effect of the additive noise (at +12 dB SNR) while the more severe lowpass conditions reflect more of the effect of the filtering. However, this explanation is inconsistent with the comparatively high correlation value of .76 between human lowpass and noise results. Instead, it may be that humans radically change their recognition strategy when dealing with severely lowpass-filtered stimuli. The loss of information on formant movements in the 1200 to 2500 Hz band may force humans to rely on a different set of cues, resulting in altered response patterns.

The final set of correlation coefficients, shown in Table 4.6, directly compares consonant pair $d'$ values between automated and human results. For the highpass condition, DST results for speech filtered with a cutoff of 1600 Hz was matched with human data with a cutoff of 1400 Hz, and DST results for a cutoff of 2800 Hz was compared with human data with the same cutoff frequency. Comparisons were made using each of the three parameterizations at both cutoff frequencies. The average correlation value of 0.50 shows little similarity between human and automated highpass error patterns given the comparable overall performance levels.

For both lowpass and noise conditions, human and machine recognition levels are not well matched. For this reason, two sets of results are presented for lowpass and for noisy speech. One set compares results at similar degradation severities, while the other tries to roughly match overall percent correct scores while keeping the degradation type the same. The lowpass comparison uses automated results for lowpass filtered speech with a cutoff of 500 Hz and human results with lowpass cutoffs of 600 Hz and 2500 Hz. Both comparisons lead to correlation values similar to the highpass comparison between the human and automated results.

Noise comparisons used automated results from speech degraded by additive noise
Table 4.6: Regression analysis values for consonant pair comparisons between human and machine recognition results. See text for details.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Mean Correlation</th>
<th>σ</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>HP</td>
<td>.50</td>
<td>.04</td>
<td>6</td>
</tr>
<tr>
<td>LP</td>
<td>.55</td>
<td>.08</td>
<td>3</td>
</tr>
<tr>
<td>600 vs 500</td>
<td>.55</td>
<td>.08</td>
<td>3</td>
</tr>
<tr>
<td>2500 vs 500</td>
<td>.50</td>
<td>.14</td>
<td>3</td>
</tr>
<tr>
<td>Noise</td>
<td>.84</td>
<td>.02</td>
<td>3</td>
</tr>
<tr>
<td>SNR -7 vs SNR 0</td>
<td>.66</td>
<td>.03</td>
<td>3</td>
</tr>
</tbody>
</table>

at 0 dB SNR and human results with noise at +2 dB SNR (roughly matching the degradations) and -7 dB SNR (matching percent correct scores). The correlation values of the comparisons show that the closest match in response patterns between automated systems and humans is seen for speech degraded by additive speech-shaped noise with matched phone recognition scores.

Figure 4-18 shows a sample comparison between human results for additive noise at +2 dB SNR and MFB results at 0 dB SNR with a .76 correlation score. Lines are placed on the graph at d' values of 3.3, corresponding to roughly 95% pairwise differentiability, and 4.7, corresponding to roughly 99.9% pairwise differentiability. Data points in the upper left section of the plot represent phone pairs for which the MFB system is making a number of errors but humans are basically perfect. The lower right quadrant (which contains no data points) corresponds to the reverse situation, with the MFB system basically perfect where humans are making errors.

There are 15 data points in the upper left quadrant of Figure 4-18. Table 4.7 lists the consonant pairs corresponding to each of the data points. For each consonant pair, it describes which of the four previously examined distinctive features have different values for the consonants making up the pair. For example, the consonant pair /p,d/ (found in the first line of the table) differs in the features voicing and place.

It is evident from the table that humans are performing much better on plosive recognition than the MFB system. 7 of the 15 consonant pairs involve two plosives,
Figure 4-18: d’ scores for MFB and human results for speech in additive noise.

and 14 out of 15 involve at least one plosive. Voicing classification errors were involved in the consonant pairs most frequently, being found for 11 pairs. Place and frication have 9 and 7 errors respectively. Sibilance errors were not found in this set of consonant pairs.

The consonant pairs shown in Table 4.7 indicate that humans are not making many of the confusions involving plosives that machines are making. These confusions most frequently involve errors in voicing classification, followed by place and frication errors. This pattern was seen across the different front ends for both the MST and DST training methods.

### 4.4 Summary of Comparison Results

Machine performance was shown to reach human levels for highpass and lowpass filtered speech when training included speech tokens of the filtering condition being tested. Information Transfer scores and Metric Multidimensional Scaling analyses reveal differences in the response patterns of humans and machines. Machines tend
Table 4.7: Description of the consonant pairs for which human performance in +2 dB additive speech-shaped noise is nearly perfect while the MFB system is making errors in 0 dB additive speech-shaped noise. An 'X' in a column headed by a phonetic feature indicates the consonant pair for that row of data differ in that feature.

<table>
<thead>
<tr>
<th>Consonant Pair</th>
<th>MFB d'</th>
<th>Human d'</th>
<th>Voicing</th>
<th>Place</th>
<th>Frication</th>
<th>Sibilance</th>
</tr>
</thead>
<tbody>
<tr>
<td>p,d</td>
<td>2.4</td>
<td>9.0</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p,g</td>
<td>2.0</td>
<td>6.4</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p,v</td>
<td>3.0</td>
<td>6.0</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>p,dh</td>
<td>3.1</td>
<td>5.3</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>t,d</td>
<td>2.9</td>
<td>10.9</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k,b</td>
<td>3.2</td>
<td>6.4</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>k,d</td>
<td>3.2</td>
<td>8.7</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>k,g</td>
<td>2.2</td>
<td>7.2</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k,th</td>
<td>2.1</td>
<td>4.9</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>b,d</td>
<td>2.5</td>
<td>6.8</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d,th</td>
<td>2.6</td>
<td>7.6</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>d,v</td>
<td>3.1</td>
<td>6.5</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>d,dh</td>
<td>2.9</td>
<td>5.7</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>g,th</td>
<td>1.8</td>
<td>5.1</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>s,x</td>
<td>3.1</td>
<td>6.7</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>11</td>
<td>9</td>
<td>7</td>
<td>0</td>
</tr>
</tbody>
</table>

to outperform humans on the sibilance and frication features. Performance on place classification is consistent with overall performance levels. For voicing, human scores are comparable to or better than machine scores.

In additive noise conditions, machine systems did not display human level performance even when the testing condition was included in the training set. Humans enjoy roughly 10 dB of SNR advantage. Information Transfer analysis reveals that the features frication and place display roughly the same gap in performance. Sibilance scores are roughly equivalent at a particular SNR despite the differences in phone recognition accuracy. Voicing scores reveal a larger gap in abilities than the gap in phone recognition accuracy. Response patterns were found to be more similar between humans and machines for additive noise conditions as compared to the filtering conditions.
Analysis of the consonant confusions for which machines made a significant number of errors while humans were nearly perfect reveal plosive classification as a weakness of the machine systems. This may indicate a weakness of HMM-based systems' ability to model the characteristics of plosive consonants, especially those characteristics relevant for voicing classification. 73% of the consonant pairs for which humans were nearly perfect but machines made significant errors involved misclassification of the voicing feature. This shows voicing classification to be a significant weakness of machine systems relative to humans.
Chapter 5

Subphonetic Features Used by Humans for Voicing Classification

The second part of this thesis explores whether knowledge of humans speech processing can be used to improve the performance of automated systems. Results of the comparison between human and machine phone classification indicate that ASR systems make the voicing distinction poorly in noise relative to humans. For this reason, voicing classification in noise (0 dB SNR) was selected as a task on which to explore ASR improvement. Classification of the initial consonant in the CVC token was chosen because these plosives are in an intervocalic position. Intervocalic plosives occur roughly twice as frequently as word-final plosives in continuous speech [Crystal & House 1988], making exploration of intervocalic plosives more relevant for continuous speech recognition systems.

In this chapter, acoustic differences between voiced and unvoiced sounds are reviewed. These cues will be termed subphonetic features as they are cues used to differentiate between a phonetic feature (voicing) being present or absent. The potential for these subphonetic features to be automatically extracted from a speech waveform at 0 dB SNR and used for voicing classification is also discussed. In the following chapters, discussion of the automatic extraction of a subset of the cues is followed by analysis of the utility of the cues in isolation and when integrated into ASR systems.
5.1 Plosive Voicing Cues

Generation of plosive sounds involves the lips or tongue creating a complete constriction in the oral cavity. During the closure interval, pressure builds up behind the constriction. When the closure is released, the pressure creates an energy burst in the speech stream and transitions in the resonances of the oral cavity.

Differences between the voiced and unvoiced plosives center around the relationship between the timing of the burst release and vocalization (i.e. the presence of a periodic source function generated by the vibrating vocal folds). Voiced consonants tend to exhibit more vocalization soon after the plosive release and in the closure interval of the plosive.

A set of subphonetic features which are identified in the literature as being potentially used for voicing classification of intervocalic plosives can be found in Table 5.1. For each feature, separate columns describe how the cue is characterized and what needs to be detected in order to evaluate a particular cue. Descriptions of each of the cues follow, with discussion of the anticipated effects of additive noise on the usefulness of the cues for voicing classification in 0 dB SNR conditions.

**Preceding Vowel Duration:** Edwards (1981) describes a study in which the characteristics of plosive consonants following a schwa vowel are examined for their utility in determining voicing classification. Measurements of the plosive characteristics were made by human operators using computers to facilitate analysis of the speech data. The durations of the schwa vowels preceding the plosives were measured and distributions determined for the schwas preceding voiced plosives and unvoiced plosives.

For schwas preceding voiced consonants, the distribution of schwa duration had a mean of 73 msec with a standard deviation of 17 ms. Schwas preceding unvoiced consonants had a mean of 65 msec with a standard deviation of 14 ms. The extensive overlap between the distributions indicated limited utility for preceding vowel duration as a voicing cue.

An examination of CVC syllables where the initial and final consonants were con-
strained to be identical revealed systematic differences in the vowel duration as a func-
tion of the voicing classification of surrounding consonants [House & Fairbanks 1953].
The vowels in the House & Fairbanks study were stressed (in contrast with the un-
stressed schwas examined by Edwards) and could be influenced by the voicing classi-
fication of the preceding or following consonant (or both). Klatt (1975) reports that
plosive voicing effects on vowel duration are strongest when the plosive is in phrase-
final position, which could be the source of the durational variations found by House &
Fairbanks using CVC tokens.

The Edwards study used speech in clean conditions. Speech degraded by additive
noise makes use of the preceding vowel duration more problematic by increasing the
difficulty of measuring the preceding vowel duration. Figure 5-1 shows spectrograms
of two sample plosives in both quiet and 0 dB noise conditions. The left column shows
an intervocalic /d/ while the right column shows an intervocalic /t/. The top row
plots show the plosives in quiet while the bottom row plots show the same plosives
in noise.

In quiet, the schwa onset and offset are easily identified in the spectrograms. In
noise, however, detection of the schwa is not as straightforward. This is in part due to
the fact that the schwa vowels in the CVC database are unstressed and in many cases
reduced. While vocalization associated with the stressed vowel is identifiable even in
noise, the schwa is much more difficult to identify (as can be seen in the figure), and
the boundaries which are clearly seen in quiet are not apparent in the noisy condition.
The decreased ability to detect the boundaries of the schwa in noise make duration
of the schwa a difficult subphonetic feature to extract reliably.

**Vocalization During the Stop Closure (Prevoicing)**: As pressure increases
behind the oral constriction before release, the transglottal pressure decreases. This
results in a cessation of vocal fold vibration when the transglottal pressure is no
longer sufficient to maintain the vocal folds’ periodic opening and closure. For un-
voiced stops, the vocal folds tend to stop vibrating more abruptly, resulting in less
vocalization during the closure period.

Prevoicing is an extension of the vocalization related to the preceding vowel. Typ-
Figure 5-1: Spectrograms of a voiced (/d/, left column) and unvoiced (/t/, right column) plosive in quiet (top row) and in noise (bottom row). The locations of the schwa onset and offset are labeled.
ically, the amplitude of vocalization during the closure is roughly 15 dB below the amplitude of the vowel [Stevens 1980]. Given that the schwa is at times difficult to detect in noise, and that any voicing found in the closure interval will be at a lesser amplitude, the detection of vocalization during the closure interval in 0 dB SNR conditions can be expected to be difficult.

Prevoicing can be seen in the quiet examples of the intervocalic plosives shown in Figure 5-1. The low-frequency energy continuing past the schwa offset boundary is prevoicing. The duration of prevoicing for the voiced plosive (/d/) is much greater than the duration for the unvoiced plosive (/t/), consistent with expectations. Examination of the noisy spectrograms shows that prevoicing is difficult to identify in 0 dB additive noise. Preliminary examinations using a range of time and frequency resolutions for spectrograms did not reveal a single spectrogram transformation which consistently revealed the presence of vocalization in the speech signal.

Edwards explored two methods of characterizing prevoicing. One method measured the absolute duration of vocalization during the closure interval. Voiced stops had a mean absolute duration of 78 msec with a standard deviation of 23 msec while unvoiced stops had a mean duration of 25 ms and a standard deviation of 15 ms. The distributions indicated an optimal criterion of 49 msec for vocalization duration, resulting in 89% correct voicing classification.

Another method involved the relative duration of vocalization during the stop closure. This measure compared the absolute duration of vocalization during the stop closure with the duration of the stop closure. The optimal criterion of 51% resulted in 92% correct voicing classification.

**Burst Amplitude:** Zue (1985) proposed using the amplitude of the burst as a cue to voicing classification. Unvoiced plosives tend to have higher burst amplitudes than their voiced counterparts with the same place of articulation due to larger pressure drops across the constriction for unvoiced plosives [Klatt 1975]. However, the differences in amplitude between plosives arising from different places of articulation is much greater than the differences between voiced and unvoiced plosives with the same place of articulation. As a result, while this cue could prove useful when the
place of articulation is known, it is difficult to implement as a general voicing cue.

Edwards reports that the distributions of amplitudes for the voiced and voiceless plosives were not different enough for this to be a useful cue.

In 0 dB SNR additive noise, detection of many bursts becomes problematic. Labial bursts (both voiced and unvoiced) are generally the weakest and are often not detectable in 0 dB SNR conditions. This makes use of burst amplitude problematic for labial voicing classification, though it remains a potential cue for differentiating between alveolar plosives.

**Voice Onset Time (VOT)**: The time period between the release of the oral closure and the onset of vocalization for the following vowel is called Voice Onset Time (VOT). VOTs tend to be shorter for voiced plosives (less than 40 ms) than for unvoiced plosives [Stevens 1980].

Figure 5-2 shows examples of the six plosives in quiet and in noise. For each plosive, one CVC token with the plosive in the initial consonant position was selected. A portion of the spectrogram including the initial consonant is shown for both quiet (first and third rows) and noisy (second and fourth rows) conditions. To determine the VOT, both the burst and the onset of vocalization following the burst need to be detected. Dashed lines on the figure show the locations of the bursts and solid lines show the locations of the voicing onsets.

In quiet conditions, the bursts and vocalization onsets are identifiable. In noise, the vocalization onsets remain identifiable but in many instances the bursts are no longer detectable. The alveolar plosives tend to be the easiest to detect, with the largest overall amplitudes and much high-frequency energy. Analysis of human confusion matrices for recognition of the CVC database in +2 dB additive speech-shaped noise reveals that alveolar plosives are recognized correctly more often than either labial or palatal plosives (96.3% for alveolars, 81.8% for labials, and 79.3% for palataals).

Edwards compared the distribution of manually-measured VOTs for voiced and unvoiced plosives. A criterion of 40 ms resulted in roughly 98% accuracy for voicing classification, making VOT an effective voicing cue in quiet.
Figure 5-2: Spectrograms of the full set of plosives in quiet and noise. The columns from left to right show labial, alveolar, and palatal places of articulation respectively. The top two rows show the voiced plosives, in quiet (top row) and in noise (second row). The bottom two rows show the unvoiced plosives in quiet (third row) and in noise (bottom row). Dashed lines show the location of the burst and solid lines show the onset of vocalization following the burst.
**Stop Consonant Duration**: Intervocalic stop consonant duration is the length of time between offset of the preceding vowel and onset of the following vowel. Stop consonant duration can be divided into the duration of the stop closure (the time between offset of the previous vowel and release of the stop) and the VOT (as described above, the time between the plosive release and onset of the following vowel).

Stop closure duration distributional differences are not great enough to make the measure useful in itself [Edwards 1981]. However, since the stop closure and the VOT both tend to be longer for unvoiced stops, stop consonant duration (being the sum of the two) has the potential to be more useful than VOT, especially given the difficulty of detecting some bursts. Edwards found that stop consonant duration could result in 85% correct voicing classification. Distributions for read speech showed a smaller amount of differentiation of voiced and voiceless stop consonant durations [Crystal & House 1988].

Detection of the offset of the preceding vowel and the onset of the following vowel are necessary for intervocalic stop consonant duration measurement. For the CVC database, the following vowel is stressed and its onset relatively easy to detect. Automatically determining the precise offset of the preceding schwa is problematic (as described above).

**Fundamental Frequency Contour**: The fundamental frequency (F0) of a vowel preceding or following a voiced plosive tends to be lower than the F0 for an unvoiced plosive ([Klatt 1975], [Haggard et al. 1981], [House & Fairbanks 1953]). Psychophysical studies have shown that rising F0 contours yield more voiced percepts than falling F0 contours ([Castleman & Diehl 1996], [Haggard et al. 1969]).

Different approaches to characterizing the F0 contour for use in voicing classification were explored by Edwards. Two of the measures compared the F0 at the vowel onset with the F0 at the offset of the preceding schwa or the average F0 over the preceding schwa and following vowel. These measures resulted in classification accuracies of 62% and 77% respectively. A third approach determined the delay between the onset of the vowel and the first local maximum in the F0 contour. This measure resulted in 77% voicing classification accuracy.
Measurement of the F0 for stressed vowels tends to be straightforward even in 0 dB additive noise. F0 measurement for schwas, however, is difficult for many of the CVC tokens due to the lower amplitude of the schwa. This makes the use of measurements requiring characterization of the F0 during the schwa problematic.

**Formant Transitions**: After releasing the burst, the articulators move to their positions for the following vowel. The longer VOTs for unvoiced plosives allow more time for the articulators to position themselves before vocalization in the following vowel begins. This results in less formant movement during the following vowel for unvoiced plosives than for voiced plosives [Stevens & Klatt 1974]. There is some evidence that the relevant perceptual cue may be formant frequencies at vowel onset rather than characteristics of the formant transitions [Summerfield & Haggard 1977].

Transitions of the second and higher formants depend upon both the position of the constriction generating the plosive and the following vowel. This creates difficulty in determining a cue based on the second formant that is applicable to the general case of plosive-vowel phone sequences.

Formant detection can be difficult even in clean conditions. In 0 dB SNR additive noise, reliable detection of the formants and their transitions is even more difficult. This can be seen by comparing the formants in clean conditions with the formants in noisy conditions in Figure 5-1.

### 5.2 Fricative Voicing Cues

Fricatives are generated by incomplete closures in the oral cavity which create a turbulent noise source. Four cues that have been identified in the literature as contributing to voicing classification by humans for fricative consonants are listed in Table 5.2. Separate columns describe the characterizations and required detections for each cue.

Each of the cues will be separately discussed along with the effects of the 0 dB SNR additive noise environment on their automatic extraction.

**Vocalization During Frication**: Different methods of analysing vocalization cues for voicing classification of fricatives have been explored. Stevens et al. (1992)
report on a series of psychophysical tests exploring the acoustic cues humans use for fricative voicing classification. They determined the primary cue was the duration of frication during which there is no vocalization. They found that when the frication period without vocalization was less than roughly 60 ms, the fricative was perceived as voiced. Analyses of syllable-initial fricatives have shown slightly longer durations for the boundary between voiced and unvoiced fricatives ([Formby et al. 1996], [Massaro & Cohen 1976]).

An alternative analysis examined the extent of vocalization at the boundaries of the frication interval [Pirello et al. 1997]. It was found that the voicing classification for a set of fricatives could be determined with 93% accuracy through analysis of an acoustic measure based on the presence or absence of vocalization in the first and last 30 msec of frication.

Both the fricative interval and glottal vibration during the fricative interval need to be detected in order to use this cue. The turbulent noise source of fricatives generates high-frequency energy. For sibilant fricatives (/s/, /sh/, and /z/ in the CVC database), this noise source dominates the spectrum. For non-sibilant phones (/th/, /dh/, and /v/), the high-frequency energy is not as great as for sibilants and therefore more difficult to detect. Detecting vocalization for both sibilant and non-sibilant phones is often difficult in speech-shaped noise because the noise is strong at low frequencies.

**Intensity and Duration of Frication**: For a read speech database, the mean duration for unvoiced fricatives was reported to be roughly twice the mean duration for voiced fricatives [Crystal & House 1988]. For a CV database, frication duration distributions were found to have a significant amount of overlap [Baum & Blumstein 1987].

Psychophysical tests using manipulated recordings of speech have shown a shift from unvoiced to voiced percept as the length of frication decreases, with a threshold around 100 msec ([Cole & Cooper 1975], [Denes 1955]).

Duration measurements require accurate detection of the fricative interval. For the stronger sibilant fricatives, accurate detection of the fricative interval boundaries is fairly straightforward. For the non-sibilant fricatives, however, the additive noise
can make accurate boundary detection difficult.

The presence of glottal vibration can add a significant amount of energy to the non-sibilant fricatives. This increased energy can be used in quiet to differentiate between voiced and unvoiced non-sibilant fricatives but is difficult to distinguish in the high energy background of 0 dB speech-shaped noise.

**Fundamental Frequency of Following Vowel:** The fundamental frequency (F0) cue for fricatives is similar to the F0 cue for plosives. As with plosives, when a vowel follows a voiced fricative it tends to begin with a lower F0 than vowels following unvoiced fricatives [House & Fairbanks 1953]. Psychophysical studies using synthetic speech shows some evidence that F0 onset may be the most perceptually relevant aspect of F0 contour [Massaro & Cohen 1976].

The effects of the additive noise degradation on the use of F0 cues are similar in the fricative case as in the plosive case.

**Formant Transitions in Following Vowel:** The vocalization that typically overlaps the frication interval for voiced fricatives causes more of the formant transition into the steady state of the following vowel to be included in vocalized speech (the beginning of the vowel). The detection of formant movements thereby becomes a cue that the particular fricative is voiced in a manner similar to the cue for plosives. Measurements of F1 transitions have shown differences in the distributions for voiced and unvoiced fricatives [Stevens et al. 1992].

As was discussed above, formant detection is a difficult problem in additive speech-shaped noise.
Table 5.1: List of potential cues to voicing classification of plosives. Separate columns describe how the cue is characterized and what needs to be detected in the acoustic signal in order to use the cue.

<table>
<thead>
<tr>
<th>Subphonetic Feature</th>
<th>Characterization</th>
<th>Detection(s) Required</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preceding Vowel Duration</td>
<td>Time between Onset and Offset</td>
<td>Onset of Preceding Vowel, Offset of Preceding Vowel</td>
</tr>
<tr>
<td>Vocalization During Closure Interval</td>
<td>Absolute or relative Duration of voicing</td>
<td>Closure Interval, Glottal Vibration</td>
</tr>
<tr>
<td>Burst Amplitude</td>
<td>Amplitude</td>
<td>Burst</td>
</tr>
<tr>
<td>Post-Burst Aspiration</td>
<td>Presence and extent of aspiration following the burst release</td>
<td>Burst, Aspiration</td>
</tr>
<tr>
<td>Voice Onset Time (VOT)</td>
<td>Time between burst and following vowel onset</td>
<td>Burst, Onset of Following Vowel</td>
</tr>
<tr>
<td>Stop Consonant Duration</td>
<td>Intervocalic period</td>
<td>Offset of Previous Vowel, Onset of Following Vowel</td>
</tr>
<tr>
<td>Fundamental Frequency (F0) Contour</td>
<td>(1) F0 at following vowel onset compared to offset of preceding vowel (2) Time to first local maximum (3) F0 at following vowel onset compared to mean during preceding and following vowels</td>
<td>F0</td>
</tr>
<tr>
<td>Formant Transitions</td>
<td>Initial formant frequencies, extents of transitions during glottal vibration</td>
<td>Formants</td>
</tr>
</tbody>
</table>
Table 5.2: List of potential cues to voicing classification of fricatives. Separate columns describe how the cue is characterized and what needs to be detected in the acoustic signal in order to use the cue.

<table>
<thead>
<tr>
<th>Subphonetic Feature</th>
<th>Characterization</th>
<th>Detection(s) Required</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocalization During Frication Interval</td>
<td>Absolute duration of unvoiced frication</td>
<td>Frication Interval, Glottal Vibration</td>
</tr>
<tr>
<td>Amplitude and Duration of Frication</td>
<td>Amplitude, duration</td>
<td>Frication Interval</td>
</tr>
<tr>
<td>Fundamental Frequency (F0) Contour</td>
<td>Extent of rise/fall</td>
<td>F0</td>
</tr>
<tr>
<td>Formant Transitions</td>
<td>Extents of transitions</td>
<td>Formants</td>
</tr>
</tbody>
</table>
Chapter 6

Subphonetic Feature Extraction

The subphonetic features associated with the voicing distinction for plosive and fricative consonants (e.g. VOT, unvoiced frication duration) occur at unknown times in the recorded waveform of each syllable. Efforts to detect automatically the acoustic events listed in Tables 5.1 and 5.2 focused on two methods of waveform analysis (see Figure 6-1). One method applied templates to a spectrographic representation of the waveform, with templates constructed to reflect attributes of the cue being detected. The second method examined the waveform using short-time autocorrelations.

In this chapter, the spectrographic template and short-time autocorrelation analysis methods will first be described. Next, the methods developed for characterizing three subphonetic features (intervocalic period, voice onset time, and F0 contour) will be discussed.

6.1 Waveform Analysis Methods

6.1.1 Spectrogram Analysis

The spectrogram-based method of analysis compared a template with a speech spectrogram to determine a measure of the similarity between the spectral pattern described by the template and the spectrogram of a speech waveform. The final output of the analysis was designed to be a frame-based parameter describing how well the
Figure 6-1: Illustration of the dual approach to exploring automatic extraction of elements of the speech signal which can be used to distinguish between voiced and unvoiced consonants.
template matched the spectrogram as a function of time using 10 msec time increments (consistent with the frame duration used in the traditional parameterizations). Templates could be heuristically generated or extracted from CVC spectrograms.

To visualize this calculation, start with a similarity measure matrix of zeros that has the same size as the CVC spectrogram (793X64). Next, imagine setting the template on top of the upper left corner of the spectrogram so that the top and left borders of the template fall along the top and left edge of the spectrogram. The convolution of the template with the spectrogram replaces the zero value in the similarity measure matrix at the center of the template location when the convolution is calculated. The template is then moved down one row relative to the spectrogram, and another convolution calculated. This convolution value replaces the zero value at the new center location of the template, one element below the previous convolution value. The template is shifted down one row and convolution values replace zero values until the bottom edge of the template meets the bottom edge of the matrix. The procedure is repeated with the template shifted one column to the right. Once convolution values have been calculated for all complete overlaps, the maximum values in each column are determined and placed in a 793-element vector. This vector is then divided into 10 msec segments (6 or 7 elements go into each frame) and the values averaged to make a 10-msec frame-based parameter.

Spectrograms were calculated using a Matlab toolkit [Slaney 1994]. Speech was first preemphasized using a two-point filter [1 -0.95]. Hamming windows of 12.8 msec length were used to extract speech segments. The square roots of the absolute values of the Fourier Transform were calculated, yielding values for 64 linearly-spaced frequency channels. Each channel contained 793 samples corresponding to 1.28 seconds of speech. The size of the spectrogram was selected based on preliminary analyses of a variety of spectrogram resolutions, providing a good tradeoff between representational ability of the subphonetic features being explored and computational complexity of spectrogram calculation and further template analyses.
A spectrogram can be treated as a matrix of spatiotemporal energy values.

\[ SpectrogramMatrix(793\times64) : S_{tf} \begin{cases} 1 \leq t \leq 793 \\ 1 \leq f \leq 64 \end{cases} \]

Templates were generated (or extracted from CVC samples) to represent a particular acoustic event to be searched for in the full set of CVC spectrograms. Examples of templates include plosive release detectors (periods of relatively low energy followed by a short burst of energy across a broad frequency range) and formant transition detectors (localized amplitude peaks varying monotonically in peak frequency).

Templates could be developed with any number of channels (up to 64) and any length (up to 793).

\[ TemplateMatrix(I \times J) : T_{ij} \begin{cases} 1 \leq i \leq I & I \leq 793 \\ 1 \leq j \leq J & J \leq 64 \end{cases} \]

The first step in template-spectrogram similarity analysis resulted in a similarity measure matrix equal in size to the original spectrogram (793X64). To produce a matrix of this size, zero values were first introduced at the edges of the similarity measure matrix\(^1\):

\[ SimilarityMeasureMatrix(793\times64) : SM_{tf} = \begin{cases} 0 & \text{if } t \leq .5I \\ 0 & \text{if } t > 793 - .5I \\ 0 & \text{if } f \leq .5J \\ 0 & \text{if } f > 64 - .5I \end{cases} \]

A measure of similarity was calculated by convolving the template and spectrogram for all complete overlaps of the template and spectrogram, producing similarity mea-

\(^1\)Note that the equations are modified when I or J is an odd integer. The modification rounds all .5I and/or .5J values down (truncating the .5 remainder) for these and all further equations.
sures to fill out the similarity measure matrix:

\[
\text{SimilarityMeasureMatrix} : SM_{t+.5f,.5j} = \sum_{j=1}^{J} \sum_{i=1}^{I} T_{ij} S_{t+i,.f+j}
\]

Conversion from the similarity measure matrix to the final frame-based output involved two steps. First a raw output vector was calculated with the same number of elements as the number of spectrogram samples (793) by taking the maximum similarity measure for each time index:

\[
\text{RawOutput}(793 \times 1) : RO_t = \max_f (SM_{tf})
\]

Next, the raw output vector was condensed to a 128-element vector by averaging over 10 msec frames to produce a frame-based similarity parameter comparable to the traditional parameterizations.

A particular template could have three different types of values: positive, negative, and irrelevant (i.e. zero). Positive values would increase the convolution value and negative values would decrease the convolution value proportionally to the (always positive) spectrogram value of whatever sample was being compared. For normalization purposes, the sum of the positive and negative values over a complete template was constrained to be zero. In addition to positive and negative values, templates could contain irrelevant elements, which would not contribute to the final correlation value. Thus, a low-frequency rising formant detector could be constructed using a set of irrelevant elements over the high-frequency range and an array of positive values surrounded by negative values in a pattern matching a rising formant.

An example of the spectrographic template approach can be found in Figure 6-2. In the figure, the top plot shows an artificially-generated onset template. This template has negative values across the frequency spectrum for 60 msec, followed by positive values across the frequency spectrum for 10 msec. The template matches spectrogram segments where a period of silence or low energy is followed by an abrupt increase in energy across the frequency spectrum, such as at plosive releases. The
Figure 6-2: An illustration of the spectrographic template approach to analyzing speech. The top plot shows a template for plosive release detection wherein a period of low energy is followed by an abrupt increase in energy across the spectrum. The middle plots show a CVC with initial consonant /t/ in quiet (second plot) and in noise (third plot). The final plot shows the frame-based onset template parameter based on matching the onset detector template with the noisy CVC spectrogram. The large peak around 200 msec corresponds to the /t/ burst.

The middle two plots in the figure show a CVC with initial consonant /t/ in quiet (the second plot) and in noise (the third plot). The final plot shows the output of the template analysis approach when the onset detector is applied to the noisy CVC.

There is a large peak in the frame-based onset detection parameter around 200 msec which coincides with the release of the initial consonant /t/. Note that the template match is maximal slightly before the burst release is found in the spectrograms. This is because the template matches a point in the spectrogram at which a period of silence is followed by an increase in energy across the spectrum. The silence-followed-by-onset is found before the 200 msec time in the spectrogram, when
the center of the onset template is still in the pre-release portion of the speech and the onset portion of the template overlaps the /t/ release.

Detection of the plosive releases is used in determining VOT. Equating the peak in the frame-based onset detection the burst release would raise difficulties if criteria for voicing classification were based on thresholds from the literature (e.g. 40 msec). This is because the method used here detects pre-burst silence, and therefore systematically shifts by roughly 20 msec the point at which the burst would be detected if the onset detection peak was used to label the burst release. Were it desired to use absolute thresholds drawn from the literature, the shift could easily be accounted for when comparing against the threshold. However, in this thesis all threshold criteria are determined after applying the methods of analysis to the CVC database and determining distributions of subphonetic features for voiced and unvoiced sounds.

Burst detection was implemented using the template method of analysis with moderate success. The alveolar plosives were most reliably detected due to alveolar plosives being strong enough to exceed the noise background level, especially in the high-frequency region. Roughly 10% of the alveolar plosive samples did not exceed detection threshold, all of which were the voiced /d/ consonant. Roughly 20% of the palatal plosives yielded no plosive detection. Labials were often not detected. The final threshold was set at a level where 30% of the labials yielded no plosive detection, but it should be noted that the threshold was a tradeoff between detecting weak plosive releases and allowing spurious plosive detection based on variations in noise. If unreasonably high VOTs (based on overall distributions of voiced and unvoiced VOTs as detected by the algorithm) are considered as non-detections, the failure rate for detection of labials rises to roughly 50%.

The template method was also applied to detection of formants in the vowels following plosives. The inherent variability of the formant transitions as well as the fact that they were commonly masked by the noise made the template approach ineffectual for the task.
6.1.2 Short-Time Autocorrelation Analysis

Smoothness Measure

The second approach to analyzing speech used short-time autocorrelation measurement of the speech waveform. One measure derived from the autocorrelation function was the number of local maxima in the autocorrelation normalized by the length of the autocorrelation function, a measure of autocorrelation function smoothness. In this measure, lower numbers correspond to smoother autocorrelation functions.

The smoothness measure was useful in that the autocorrelation function during vocalization was generally smooth (few local maxima, usually associated with harmonics of the fundamental frequency and the formants), while during sibilant and aspirated regions the autocorrelation was very rough, with a large number of local maxima. The resultant measure is low during vocalization, high in sibilant/aspirated regions, and medium in the noise background.

For this type of analysis, an autocorrelation was calculated every 10 msec using a 20 msec window (200 speech samples). Figure 6-3 shows an example of the autocorrelation smoothness measure for a noisy CVC token. The CVC contains two sibilant consonants, an initial /z/ and a final /sh/. The sibilant regions of speech show sustained high values of the smoothness measure (indicating rough or non-smooth autocorrelations for these regions) sustained throughout the sibilant region of speech. The sibilant sounds differed from each other somewhat, with /z/ generally producing the highest values. Methods for using the smoothness values to differentiate between the sibilant sounds were explored but never refined to the point where they could be used for voicing classification. Post-burst aspiration is also detected through use of the smoothness measure.

The two vowel portions of the CVC, the preceding schwa and the CVC vowel, both show low smoothness measures in the figure. Initial efforts at extracting subphonetic features utilized the smoothness measure to detect vowels with some success, but the method tended to fail for a large number of schwas. This seems due in part to the low amplitude and short duration of the schwa, which did not cause the autocorre-
Figure 6-3: Example of the short-time autocorrelation smoothness measure. The top plot shows the sampled CVC with initial consonant /z/ and final consonant /sh/. The middle plot shows the same CVC in additive noise at 0 dB SNR. The bottom plot displays the smoothness measure of the short-time (20 msec) autocorrelation for the noisy CVC token. Regions of vocalization (roughly 125 to 175 msec and 300 to 400 msec) are characterized by low values (less than .1) of the smoothness measure (indicating a smooth autocorrelation). The short peak from roughly 200 to 250 msec corresponds to the initial /z/ consonant, while the plateau extending from 500 to 600 msec corresponds to the final /sh/ consonant. Both of these consonants are sibilant, which typically result in high values of the short-time autocorrelation measure.
lation measure to remain at low values for long enough periods to differentiate from variability in the smoothness measure generated by the additive noise.

**F0 Detection and Estimation**

Short-time (25 msec) autocorrelations were also utilized for fundamental frequency (F0) detection and estimation. The peak in the autocorrelation in the range of delays that correspond to the normal F0 range (80-250 Hz) was determined. This peak value was compared to a threshold based on the short-time energy as calculated by the autocorrelation with no delay. If the peak value was greater than the threshold, then F0 was considered to be detected, corresponding to vocalization in the speech signal. A threshold of $0.25 \times \text{energy}$ was found to be most effective in compromising between the number of false vocalization detections and false vocalization rejections. F0 detection was used as part of the vowel detection process.

If F0 detection was successful (the autocorrelation peak was above the threshold), F0 was estimated as $1/(\text{peak autocorrelation delay})$. The conversion from peak autocorrelation delay to F0 was often done for purposes of ensuring reasonable values were being produced. However, the main use of F0 estimation, namely its analysis at the onset of vowels following the consonant being classified, did not convert to specific fundamental frequency, instead dealing with peak autocorrelation delay directly.

### 6.2 Subphonetic Feature Detection

The spectrographic and autocorrelation measurements were combined in an attempt to improve detection of subphonetic features. Although many features were studied (including all of those listed in Tables 5.1 and 5.2), only three could be characterized reliably at 0 dB SNR; intervocalic period (IVP; the time between the schwa offset and the vowel onset), Voice Onset Time (VOT; the time between the release of a plosive and the onset of the following vowel), and delta fundamental ($\delta F0$; the change in F0 at the onset of the vowel). For all of these cues, the extraction approach was based on detection of a vowel onset landmark and targeted analysis based on the location
of the vowel onset. After a brief discussion of landmark detection, each of the three subphonetic cues found useful for voicing classification in 0 dB noise will be discussed.

6.2.1 Landmark Detection

The most consistently detectable acoustic event across the CVC tokens was the onset of the stressed vowel following the initial consonant. Vowel onsets were detected through analysis of the short-time autocorrelation function, as discussed in Section 6.1.2. Determination of the location of the vowel onset aided detection of other acoustic events (e.g. plosive release) by determining a reasonable time window over which the other acoustic events were likely to be detected.

A vowel was considered detected when vocalization was detected (through comparison of the autocorrelation peak over a range of delays with a threshold based on the short-time energy) for at least 3 consecutive frames (30 msec) while the short-time energy of the speech signal was above its mean value over the entire CVC token. This detection criterion was met by all stressed vowels.

This method of detecting vowel onsets resulted in better detection scores than those reported for a system examining energy levels in six different frequency bands [Liu 1995]. It should, however, be noted that Liu’s system was developed on speech in quiet and tested on speech in noise.

To measure how specific to the 0 dB noise condition the vowel onset landmark detection approach was, a set of tests was performed applying the algorithm to speech at other Signal to Noise Ratios. The algorithm used in 0 dB conditions proved fairly robust for SNRs ranging from -5 dB to +10 dB. At -10 dB the detection of vocalization largely failed. In higher SNR conditions, the thresholds produced many spurious vocalization detections, generating a large number of false landmarks.

6.2.2 Intervocalic Period

Measurement of the InterVocalic Period (IVP, see Figure 6-4) involved determining the interval between the onset of a vowel and the offset of any vocalization preceding
the vowel. When the IVP is less than 100 msec, the initial consonant is very unlikely to be unvoiced.

The criterion for the detection of stressed vowels was described above. There were a number of cases where the unstressed schwa did not meet this criterion. Furthermore, many times when vocalization was detected, the vocalization detection threshold was not met near the ends of the vocalization. For these reasons, additional mechanisms were used to determine regions of vocalization.

First, the energy at the offset of a period of detected vocalization was compared to the energy contour of the CVC. Vocalized speech segments are relatively high in energy. When vocalization was detected, the region labeled as vocalized was continued past the end of its original detection as long as the energy level of the CVC remained high (more than 1.2* average energy value over the CVC).

When a vowel was detected but no preceding vocalized segment was found in a window extending from 300 msec to 50 msec before the vowel onset through autocorrelation smoothness analysis alone, a second pass was made with a relaxed set of criteria. Instead of requiring 3 consecutive frames that met the vocalization detection criterion, only one frame of vocalization detection was required as long as the energy contour around the vocalization was high (using the same threshold of 1.2* mean energy over the CVC). If this second pass did not reveal the schwa location, the criteria were relaxed even further, requiring a short-term high-energy segment without any F0 detection. Inclusion of these stages allowed for detection of over 95% of the schwas in the database. An example of intervocal period determination is shown in Figure 6-4.

The IVP was defined as the number of frames between the onset of a vocalized region and the offset of any previous vocalized region (up to a maximum of 30 frames). The output of IVP analysis was a frame-based parameter which was set to zero for all frames except those determined to be in the intervocalic period. All frames in the intervocalic period were set to a value equal to the total number of frames in the IVP (as seen in Figure 6-4).

The distributions of intervocalic periods for voiced and unvoiced initial consonants
Figure 6-4: Example of IVP detection. The top plot shows a CVC in quiet, the second plot the same CVC in 0 dB additive noise. The third plot displays the results of the vocalization detection algorithm applied to the 0 dB CVC. The stressed vowel is detected by the strictest criterion, and its onset is determined as the first of three consecutive frames meeting the vocalization detection criterion in which the energy value exceeds $1.2 \times$ mean average. In the same plot is included the energy contour (the dotted line) along with the average energy value (the dashed line). The final plot shows the frame-based IVP as determined by the information presented in the third plot, uniformly set to the total detected IVP from the offset of the previous vocalization to the onset of the CVC vowel.
Figure 6-5: Distribution of IVP for unvoiced (top plot) and voiced (bottom plot) initial consonants in 0 dB additive noise. An IVP of 0 indicate failure to detect vocalization prior to the onset of the vowel.

in 0 dB additive noise are shown in Figure 6-5.

6.2.3 Voice Onset Time

Determination of VOT depends upon two detections: the point in the speech at which the constriction is released (the burst), and the onset of the following vowel. The vowel onset was detected using the procedure described for IVP detection in Section 6.2.2. Upon detecting a vowel onset, plosive release detection was initiated over a window of time before the vowel onset.

Detection of the set of plosive bursts is difficult due to their variabilities. /t/ bursts are most distinctive, including an abrupt increase in energy across much of the high-frequency spectrum, and a period of aspiration following the burst. /b/ bursts are the most difficult to detect because the energy present at the burst is relatively small (compared to /t/ bursts), and because this energy is focused in the low-frequency portion of the spectrum (where the additive speech-shaped noise has
most of its energy).

Many different templates, both artificially generated and extracted from speech tokens, were tested for their ability to consistently result in detection of bursts in 0 dB additive noise. A simple onset template, shown in Figure 6-2, modeling a period of relatively low energy (corresponding to the time during which the constriction in the vocal tract builds up pressure in the oral cavity) followed by an abrupt increase in energy across the frequency spectrum was found to be as effective over the full plosive set as a number of complex artificial burst spectra or extracted burst spectra.

Each time a vowel was detected, a window of time before the vowel onset was analyzed. This window extended from 20 msec after the end of the previous vocalic segment (corresponding to the schwa) to 40 msec before the vowel onset. In those cases where the preceding schwa was not detected, the window began 300 msec before the onset of the vowel.

Detection of a burst within the window proceeded in a number of stages. Each stage had a slightly different criterion, with the criterion of each stage being less restrictive than the previous stage’s criterion. Analysis stopped once a burst was detected.

The first stage of analysis looked for distinctive burst energy (as determined by the artificial burst template matching the CVC spectrogram) followed by high values of the short-time autocorrelation smoothness measure (corresponding to post-burst aspiration). These characteristics are typical for the /t/ and to a lesser extent the /d/ bursts. In addition, they characterize the onset of sibilant consonants (e.g. /s/), with the main difference being that sibilant regions tend to maintain high values of the autocorrelation smoothness measure for longer durations. The distributions of times between sibilant onset and vowel onset for voiced and unvoiced sibilant consonants overlapped largely with the distributions for the voiced and unvoiced plosives respectively. This made the VOT feature useful in discriminating between sibilant sounds as well as plosives.

The following analysis stages reduced the thresholds for the match between the CVC spectrogram and the artificial burst template, and of thresholds for noisiness
in the autocorrelation for detecting aspiration. The thresholds are lowered in successive stages. Eventually the burst energy threshold is removed from the detection requirement, meaning release is detected based only on the ability to detect post-burst aspiration. If no burst was detected using the thresholds, a further analysis stage examined the energy contour at the vowel onset. In those cases where the burst release is within 20 msec of the vowel onset, detection of the burst release using template matching is prevented by the onset detector’s match with the vowel onset. Characteristically, the energy contour in these cases displays two peaks in close succession. For those instances where no plosive was detected using template matching over a window ending 4 frames before the vowel onset, if the energy contour displayed two peaks in close succession a value of 2.5 frames for the VOT was used as the output. Roughly 13% of the voiced plosives were detected in this manner while only around 3% of the unvoiced plosives met these criterion. No analytical methods were able to detect the full set of plosives, with /b/ and /p/ burst detection being particularly difficult.

For all CVCs, when a vowel onset was detected, the burst detection method described above was performed. In the case where a burst was detected, the time (in frames) between the burst location and the vowel onset location was calculated and used as the VOT. In a manner similar to IVP, the VOT parameter was set to zero for all frames except those falling between the burst (or sibilance onset) and onset of the following vowel, which were uniformly set to the VOT.

Figure 6-6 shows histograms for VOT values for the set of voiced and unvoiced initial consonants in 0 dB additive noise. For a large number of consonants no burst was detected. This is reasonable since the measure was designed for plosive classification but used for both plosives and fricatives. Histograms for VOT values for the plosive consonants are shown in Figure 6-7. Using a criterion of 8 frames for distinguishing between voiced and unvoiced plosives, 88% correct voicing classification can be realized for plosives in 0 dB SNR additive noise, compared to Edwards’ finding of 98% correct for plosives in quiet.
Figure 6-6: Distribution of VOT values for unvoiced (top plot) and voiced (bottom plot) initial consonants. Note that a VOT value of 0 is returned when no burst is detected prior to the onset of a vowel. VOT values of 2.5 frames indicate failure of template matching to detect a burst release but detection of a double peak in the energy contour at the vowel onset characteristic of voiced plosives. There were roughly 200 instances where no burst was detected for voiced initial consonants and around 100 instances for unvoiced initial consonants.
Figure 6-7: Distribution of VOT values for voiced (top plot) and unvoiced (middle plot) plosives. The bottom plot shows percent correct voicing classification as a function of different VOT criteria.

6.2.4 Delta Fundamental

The final parameter extracted from the CVC tokens was a measure of the slope of F0 at the onset of vowels, termed Delta Fundamental (δF0). This measure was calculated as the average first difference between sample delays that resulted in maxima in the short-time autocorrelation (i.e. the sample delays were not converted to F0) over the first 9 frames (90 msec) of the vowel. The 9 frame length of analysis was selected based on preliminary examinations of the utility of the measure using a number of different frame lengths.

To avoid unreasonably large Delta Fundamental values when F0 was poorly estimated (for example when the estimated F0 jumped from 100 Hz at the first frame of the vowel to around 200 Hz for the remaining frames in the vowel), a maximum absolute value for the first difference of roughly 10% of the estimated F0 was introduced. Any first difference which exceeded this maximum rate of change was removed from the Delta Fundamental calculation. Spurious values were common in the initial
Figure 6-8: Distribution of Delta Fundamental values for unvoiced (top plot) and voiced (bottom plot) initial consonants.

frames of vowels, being present in roughly a third of the CVC tokens.

Histograms of the distributions of $\delta F_0$ values for voiced and unvoiced initial consonants can be found in Figure 6-8. Because the values used in calculating $\delta F_0$ were sample delays and not frequencies, a positive average first difference actually corresponds to a negative slope of F0. The distribution for voiced initial consonants is centered around 0. For unvoiced initial consonants, the distribution is shifted to positive values. This indicates that F0 generally decreases for unvoiced sounds while being equally likely to rise or fall for voiced sounds.
Chapter 7

Subphonetic Feature Extraction

Results

7.1 Voicing Classification Scores Using Subphonetic Features

Voicing classification of the initial CVC consonants was performed using a nearest neighbor algorithm in the subphonetic feature space. The CVC database was first divided into two sets of alternate tokens from an alphabetized list. A Euclidean distance was computed between the subphonetic feature parameters of each CVC token from the even set and all CVC tokens from the odd set to determine which odd token's subphonetic features were closest in subphonetic feature space. Comparisons were also made between tokens from the odd set with the set of even tokens. The voicing category of the initial consonant for the closest CVC match was taken as the voicing classification output. Raw values of VOT and IVP (in frames) were used, while δF0 values were scaled by a factor of 10 in order to roughly equalize weightings for the different subphonetic features.

Table 7.1 shows the results of the subphonetic features when used for classifying initial consonants as voiced or unvoiced. Humans achieve roughly 95% correct voicing classification in 0 dB SNR additive noise. Scores for the traditional systems range
from 80-85 percent correct.

Table 7.1: Percent correct voiced/unvoiced categorization based on various combinations of the subphonetic features and comparison values from human and traditional ASR results. The Human score is in parentheses because it is an interpolated score using voicing classification scores in +2 and -4 dB SNR additive noise. VOT2 is a simplified VOT detector using only the strictest criterion for burst detection during VOT estimation.

<table>
<thead>
<tr>
<th></th>
<th>System/Parameters</th>
<th>Percent Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>Human</td>
<td>(95)</td>
</tr>
<tr>
<td>Traditional Systems</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MFB</td>
<td>85</td>
</tr>
<tr>
<td></td>
<td>CEP</td>
<td>84</td>
</tr>
<tr>
<td></td>
<td>EIH</td>
<td>80</td>
</tr>
<tr>
<td>Subphonetic Feature Pairs</td>
<td>IVP VOT</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>VOT</td>
<td></td>
</tr>
<tr>
<td></td>
<td>δF0</td>
<td>78</td>
</tr>
<tr>
<td></td>
<td>IVP</td>
<td></td>
</tr>
<tr>
<td></td>
<td>δF0</td>
<td>87</td>
</tr>
<tr>
<td>Subphonetic Feature Triples</td>
<td>IVP VOT</td>
<td>88</td>
</tr>
<tr>
<td></td>
<td>δF0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>IVP</td>
<td></td>
</tr>
<tr>
<td></td>
<td>δF0 VOT</td>
<td>89</td>
</tr>
</tbody>
</table>

Voicing classification accuracy using pairs of subphonetic features ranged from 76-87 percent. Note that use of two subphonetic features, IVP and δF0, achieved 87% correct voicing category determination. This is comparable to the voicing score for the best of the traditional ASR approaches (85% for the MFB system) and comparable to results for a Cepstral/HMM system trained to make the voiced/unvoiced distinction for clean speech [Eide 1993] (albeit with a larger set of phonemes).

When all three subphonetic features are used, 88% correct voicing classification is achieved. The inclusion of VOT with IVP and δF0 improved performance by only 1%. A test was also performed using an altered VOT detector. This detector, VOT2, included only the strictest burst detection criterion (requiring onset detection followed
by aspiration in the burst window before the vowel onset). VOT2 together with IVP and δF0 resulted in the best score: 89% correct.

7.2 Use of IVP and Delta Fundamental in Quiet

The strength of just two subphonetic cues in the 0 dB noise condition, IVP and δF0, motivates further research on the use of multiple cues in quiet. A preliminary analysis of the two cues for the clean CVC database was performed to determine the voicing classification ability of the two cues for undegraded speech. Waves+, a signal processing toolkit (Entropic Research Laboratory, Inc.) was used to analyse clean CVC tokens for energy countours and F0. Upon determining the energy and F0 contours, the algorithms developed for noisy speech were used to determine vowel onset landmarks and then characterize the IVP and δF0 subphonetic features. Nearest neighbor analysis of the subphonetic feature parameters was then used to determine voicing classification scores.

The two features yielded voicing classification scores of 87% for initial consonants in CVCs degraded by 0 dB additive noise and 91% for clean speech. This shows that the IVP and δF0 cues are relatively robust. The voicing classification score decreases by only 4% between clean speech and speech degraded by 0 dB SNR additive noise.

7.3 Orthogonality of Subphonetic Features and Traditional Systems

The next stage in exploring the utility of subphonetic features for improving ASR systems examined how similar traditional system response patterns were to subphonetic feature response patterns. Similarity was measured by determining voicing classification for individual CVC tokens by traditional ASR approaches and the subphonetic feature extraction approach. To the extent that voicing classification based on subphonetic features makes errors on different tokens than an ASR system, combination of the two may result in improved accuracy.
If two systems achieved 100% correct voicing classification, the responses of the two systems for each individual CVC token would necessarily be identical. If two systems each achieved 90% correct voicing classification, the percentage of CVC tokens labeled with the same voicing classification by the two systems would have to range between 80% and 100%. If 100% of the CVC tokens were labeled identically, then the response patterns of the two systems would be identical and the potential for improving the traditional ASR system through incorporating subphonetic feature extraction would be minimal. If 80% of the CVC tokens were classified identically, then the systems would be maximally orthogonal. Every CVC token which the traditional system mislabeled would have to be labeled correctly by the subphonetic feature system, and all tokens mislabeled by the subphonetic feature system would have to be labeled correctly by the traditional system. This condition indicates the best opportunity for improving ASR performance through integration with subphonetic feature extraction.

Table 7.2 shows the percentage of tokens that were labelled with different voicing categories for each pair of approaches. Included in the table are the possible ranges over which the percentage of different labels could vary given the percent correct scores of the systems being compared. Comparisons of the FTR2 system (the subphonetic feature extraction approach using IVP and δF0) with the traditional ASR systems show significantly different error patterns. This indicates that the CVCs misclassified by traditional ASR systems tend to be classified correctly with the subphonetic feature approach, and vice versa. This situation indicates potential for improving ASR systems through integration with subphonetic feature extraction.

### 7.4 Integration of Subphonetic Feature Extraction with Traditional Parameterizations

The final set of experiments explored whether the subphonetic feature extraction approach could be integrated with traditional parameterizations to increase the accuracy of voicing classification. The subphonetic features were appended to the traditional
Table 7.2: Percentage of initial consonant tokens which were labelled with different voiced/unvoiced category by pairs of approaches. In parentheses are the ranges over which the percentage could vary given overall percent correct scores. FTR2 is the subphonetic feature extraction approach determining IVP and Delta Fundamental.

<table>
<thead>
<tr>
<th></th>
<th>MFB</th>
<th>CEP</th>
<th>EIH</th>
</tr>
</thead>
<tbody>
<tr>
<td>FTR2</td>
<td>24 (2-28)</td>
<td>25 (3-29)</td>
<td>26 (7-33)</td>
</tr>
<tr>
<td>MFB</td>
<td>18 (1-29)</td>
<td></td>
<td>25 (5-35)</td>
</tr>
<tr>
<td>CEP</td>
<td></td>
<td>24 (4-36)</td>
<td></td>
</tr>
</tbody>
</table>

parameter vectors for each frame. This was a straightforward matter since the output of the subphonetic feature extraction method was designed to be a frame-based parameter with one value per 10 msec frame.

HMM systems with the same numbers of states and mixtures as those already tested were trained using Degradation Specific Training. As can be seen in Table 7.3, appending the subphonetic features increased phone classification scores for all but one case. In this case, MFBs appended with IVP and delta F0, frication and place scores decreased. Improvements were always seen in the accuracy of voicing classification. For the MFB and CEP systems, the improvements made up roughly half of the difference between human and machine voicing scores, and roughly a third for the EIH system.

Figure 7-1 shows voicing scores for each of the consonants. The traditional systems displayed relatively poor voicing scores for the voiced plosives. Average voicing scores across the automated systems were 91% and 64% correct for unvoiced and voiced plosives respectively. The asymmetry in voicing scores for the plosives has been reported in two previous studies using HMMs. Ghitza's study [Ghitza 1993] showed the asymmetry in noise ranging from +30 to +10 dB SNR for both a Fourier-based spectrum and the EIH. Work by Duchnowski (1993) explored the use of multiple HMM recognizers separately recognizing band-limited speech and found a 29% voicing classification error rate for voiced plosives and a 9% error rate for unvoiced plosives. The source of this recognition asymmetry is unclear.
Table 7.3: Performance of traditional systems before and after inclusion of subphonetic features. Results are percent correct values for phone classification (data columns 1, 3, and 5) or voicing classification (data columns 2, 4, and 6).

<table>
<thead>
<tr>
<th>ASR System Parameters</th>
<th>Percent Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Traditional Parameterization</td>
</tr>
<tr>
<td></td>
<td>Alone</td>
</tr>
<tr>
<td>Phone</td>
<td>Voicing</td>
</tr>
<tr>
<td>MFB</td>
<td>53</td>
</tr>
<tr>
<td>CEP</td>
<td>50</td>
</tr>
<tr>
<td>EIH</td>
<td>41</td>
</tr>
</tbody>
</table>

Incorporating subphonetic features improves voicing scores for the MFB and cepstral parameterization primarily for the voiced plosive consonants. For the EIH parameterization, incorporation of subphonetic features improved voicing scores primarily for unvoiced plosives.

7.5 Summary of Subphonetic Feature Extraction Results

Only three of the subphonetic features examined were automatically extractable from speech in 0 dB SNR additive noise conditions; IVP, VOT, and δF0. None of these features alone provided more than 68% correct voicing classification ability.

Taken together, IVP and δF0 yield voicing scores of 87% correct. Addition of VOT only improves voicing accuracy by an additional one or two percent.

Analysis of the CVC tokens for which the traditional systems were making errors and for which subphonetic feature extraction lead to voicing misclassifications showed potential benefits for hybrid systems. To test the benefit of incorporating subphonetic feature extraction with traditional systems, frame-based parameter vectors were augmented with the output of the subphonetic feature extraction system. While
Figure 7-1: Percent correct voicing categorization for the traditional approaches with and without the subphonetic features. The top two plots, showing results for MFB and Cepstral parameterizations, show improvement in the phones /b/, /d/, and /g/. The bottom plot, showing results for the EIH parameterization, does not show improvement for these phones although voicing classification is still improved. See text for further discussion.
phone classification scores were not always improved (in one case frication and place scores degraded more than the voice scores improved), voicing classification scores were consistently improved. A third (for EIH) to a half (for MFB and MFCC) of the differences in voicing classification accuracy between the traditional systems and humans can be made up for by incorporating subphonetic features.
Chapter 8

Discussion

8.1 Human and Machine Performance Comparison

Examination of phone classification error rates shows that machines can perform at human levels on speech that is either highpass or lowpass filtered provided they are trained on the test condition. When the test condition was included in the training set, either using the MST or DST approach, results were far superior to results when the test condition was not included in the training set. Differences between MST and DST results were small.

Analysis of confusion matrices shows differences in the error patterns of humans and machines. In highpass and lowpass conditions, machines make fewer frication and sibilance classification errors while humans make fewer voicing classification errors. Place classification scores are comparable. The differences in response patterns suggest that humans and ASR systems rely on different properties of the acoustic speech waveform for recognition despite the overall performance level similarities.

Machine response patterns were found to be similar for all three front ends tested. In particular, the use of Ghitza's Ensemble Interval Histogram (EIH) auditory-based front end did not significantly alter the overall performance levels or the error patterns.

In additive noise conditions, humans outperformed the ASR systems. Human
and machine error patterns were more similar than in filtering conditions, especially when overall performance levels were matched, though not as similar as the response patterns of the different ASR systems. As in the filtered conditions, machines displayed relatively good performance for sibilance classification and relatively poor performance for voicing classification.

As the effectiveness of the speech cues is decreased through filtering, the gap between human and machine abilities decreases. This differentiates the effects of filtering from those of additive noise. In the case of additive noise, the gap in recognition accuracy between humans and machines increases as available information decreases.

The human results show that there are enough reliable cues in the speech signal degraded by additive noise to allow for performance at levels above those displayed by the ASR systems. Error pattern analysis points to voicing classification as a particular weakness of the ASR systems relative to humans. This motivated use of voicing classification in 0 dB additive noise as a task on which to explore the utility of using human speech processing knowledge to improve ASR systems.

8.2 Subphonetic Feature Analysis

A number of studies have explored the perceptual contributions of elements of the acoustic signal to voicing classification of consonants. These studies use signal processing methods to manipulate cues from recorded speech and systematic variations of cues in synthetic speech. A number of cues that are perceptually relevant for voicing classification of consonants have been identified for clean speech.

Automatic detection and characterization of the signal elements shown to be perceptually significant for humans provides a kind of front end reflecting knowledge of human processing. This method of analysis has been termed subphonetic feature extraction, and is an extension of previous efforts to find invariant cues for phonetic features.

The subphonetic feature extraction model implemented in this thesis used vowel onsets as landmarks in the speech signal. In the landmark approach, it as assumed
that information is not distributed equally over the duration of the speech signal, and that detection of landmarks can indicate points in speech where phonetically relevant information is present. Detection of landmarks was based on detection of vocalization and on signal energy analysis. Upon detecting a vowel onset, targeted analysis looked for acoustic cues to voicing over specific time windows relative to the landmark. This approach is fundamentally different from the HMM approach, which generates a number of models for phones and then attempts to determine the most likely set of phone models to have produced an acoustic signal. HMMs use the same set of parameters for every frame in the speech signal, and every frame contributes equally to the labelling of speech.

Two of the cues found useful for voicing classification in 0 dB noise were temporal in nature, involving determination of intervocalic period and voicing onset time. The distributions for these cues were shown not to be exponential, but rather resembled gaussian distributions (see, for example, Figure 6-7).

Standard HMMs can model temporal cues in two ways. One way is through incorporation of temporal information in the parameter set. This is typically done through use of δ-coefficients, which measure rates of change over 20 to 50 msec intervals. The second method of modeling duration is through state transitions. For example, if the first state of a three-state HMM plosive model represents the closure interval, the second state represents the time from the burst to the onset of vocalization, and the third state represents the vocalization, then VOT (the time between exiting the first state and entering the third) is modeled through the number of self-transitions in the second state. The probability distribution as a function of the length of time in the second state is exponential, being the probability of self-transition taken to the power of the number of self-transitions. Standard HMMs are therefore ill-suited to modeling distributions of VOT to differentiate between voiced and unvoiced plosives. Modifications to the standard HMM approach have sought to introduce non-exponential distributions into the HMM framework [Rabiner & Juang 1993]. It is possible that these non-standard HMMs using more appropriate duration models would be able to accurately model these cues.
VOT was found to be useful for voicing classification for subsets of the consonants. The plosives /t/ and /d/ were nearly perfectly classified (98% correct). Application of this classifier to the entire plosive set resulted in 88% accuracy. When applied to the entire consonant set, the VOT measure is much less useful.

IVP was a useful cue for both plosives and fricatives. For both types of consonants, finding a short consonant could rule out the consonant being unvoiced.

The final cue for consonant voicing classification involved the F0 contour during the first 90 msec of the vowel following the consonant. For this cue, the information relevant to consonant classification is found in the portion of the speech signal which the HMM typically assigns to a vowel model. The independence assumption for state output probabilities prevents the F0 contour from being used for consonant classification by a standard HMM using phone models. HMMs which use bigram or trigram models could potentially mitigate this limitation.

The integrated HMM and subphonetic feature extraction approach solves this problem by characterizing the F0 contour and including the resulting parameter with the speech parameterization over five frames prior to the vowel onset. This makes the information available within the consonant segment and therefore available for consonant classification.

Consistent with the results of previous work evaluating voicing cues in quiet, no single cue classified the plosive or fricative consonant sets perfectly in noise. The use of multiple subphonetic features increased voicing classification performance significantly. This indicates that the chosen cues were not redundant. The use of just two of the subphonetic features, IVP and δF0 allows voicing classification at levels comparable to the traditional ASR systems. While voicing classification scores for the subphonetic features exceeded scores for traditional ASR systems, human levels of performance were not achieved.

The subphonetic feature extraction approach was shown to classify correctly a large number of the CVC tokens that the HMM systems misclassified. This shows that HMMs are using a different set of cues, and motivates attempting to improve HMM performance by supplementing their input with subphonetic features. The
integrated HMM and subphonetic feature system decreased the gap between machine
and human voicing classification performance by a third for the EIH parameterization
(from 20% errors to 14%, compared with 5% for humans) and by roughly half for the
MFB (15% to 8%) and the MFCC (16% to 9%) parameterizations. The classification
accuracy of the subphonetic features alone and integrated into the HMM framework
motivates further work on the extraction of subphonetic features for use in improving
the ability of ASRs.

8.3 Future Work

The work reported in this thesis explored the use of subphonetic features to improve
voicing classification in 0 dB additive noise. Voicing classification was shown to
be a relative weakness of machine systems in the filtering conditions as well as the
additive noise condition. Investigation of the utility of the approach for improving
voicing classification of filtered speech would therefore be of interest. Many of the
cues to voicing that could not be used in 0 dB SNR conditions may be useful in less
severely-degraded speech. Automatic extraction of these other cues would allow for
objective measurements of their utility in the voicing classification task.

Vowel onsets are not the only landmarks that can be used for locating areas in the
speech stream that contain acoustic cues. For example, vowel offset landmarks can
identify another portion of speech over which F0 analysis could prove beneficial to
csonsonant classification. Expansion of the subphonetic feature extraction approach
to include detection and use of other types of landmarks could prove beneficial for
voicing classification and classification of other phonetic features (or even applying
the subphonetic feature approach to phone classification directly).

The use of subphonetic features easily extends to classification of other phonetic
features. For example, high-frequency energy detection could be used to detect sibi-
lant phones, or characteristics of a burst could be used to determine the place feature
of a particular plosive. A larger number of features could be analysed simultaneously
or in a multi-stage (possibly hierarchical) manner such that examination of a sub-
set of the subphonetic features determines which other subphonetic features will be analysed further.

The performance of the subphonetic feature extraction approach on a variety of speech corpora also bears examination. The database used in this research included isolated two-syllable speech tokens produced by only two speakers. This provided a very constrained set of speech on which the machine systems were trained and tested. Testing on a corpus including more speakers would be necessary to determine how effective subphonetic features are for speaker-independent recognition.

Moving from isolated speech tokens to continuous speech would also introduce many new issues for the use of a subphonetic feature extraction system. The distributions of subphonetic features are likely to be altered in a continuous speech corpus. In addition, the presence of consonant clusters complicates the choice of time windows over which to examine speech for subphonetic features when a landmark is detected.

Another issue that bears examination is alternative methods of utilizing automatically extracted subphonetic features. Two methods were tested in this thesis. The first performed a nearest neighbor analysis using a Euclidean distance metric in the subphonetic feature space. Hierarchical analysis might use extracted subphonetic features more effectively. Initially, features useful for determining broad classifications of phones could be analysed, determining for example whether a phone was a plosive or a fricative. Further stages of analysis would examine those subphonetic features relevant for classifying phones within the broad phonetic class.

In this thesis, extracted subphonetic features were integrated with traditional ASR systems through supplementing traditional parameterizations and using HMMs to classify the augmented parameter set. An alternative approach would be to use the subphonetic features with the output of HMM analysis of traditional parameterizations. For example, HMMs could be used to classify the broad phonetic category and determine relevant subphonetic features for further classification. Alternatively, pattern recognition algorithms (e.g. artificial neural networks) could be used to perform phone classification based on HMM classification and a set of subphonetic features.
Appendix A

Information Transfer Scores for MFCC and EIH Systems

This appendix contains figures analogous to Figures 4-13 through 4-15 showing Information Transfer (IT) scores for the MFCC and the EIH system for each of the three degradations tested.
Figure A-1: Information Transfer (IT) scores from human CVC and CV data and from MFCC systems on highpass filtered speech.
Figure A-2: Information Transfer (IT) scores from human CVC and CV data and from MFCC systems on lowpass filtered speech.
MFCC/Human Noise Comparison: IT Scores

Figure A-3: Information Transfer (IT) scores from human CVC and CV data and from MFCC systems on noisy speech.
Figure A-4: Information Transfer (IT) scores from human CVC and CV data and from EIH systems on highpass filtered speech.
EIH/Human Lowpass Comparison: IT Scores

Figure A-5: Information Transfer (IT) scores from human CVC and CV data and from EIH systems on lowpass filtered speech.
Figure A-6: Information Transfer (IT) scores from human CVC and CV data and from EIH systems on noisy speech.
Bibliography


THESIS PROCESSING SLIP

FIXED FIELD: ill. name

index biblio

▷ COPIES: Archives Aero Dewey Eng Hum
Lindgren Music Rotch Science

TITLE VARIES: □

NAME VARIES: □ Jonathan

IMPRINT: (COPYRIGHT)

▷ COLLATION: 1114

▷ ADD: DEGREE: □ □ □ □ DEPT.:

SUPERVISORS:

NOTES: Fiche goes to science

cat#: date: F 76

▷ DEPT: □ □ □
▷ YEAR: □ □ □ □ □ □ □ □ DEGREE: □ □ □ □ □ □ □ □ □ □ □ □
▷ NAME: □ □ □ □ □ □ □ □ □ □