Automatic Time Alignment of Phonemes Using Acoustic-Phonetic Information

John-Paul Hosom
B.S., University of Massachusetts at Amherst, 1987

A dissertation submitted to the faculty of the
Oregon Graduate Institute of Science and Technology
in partial fulfillment of the
requirements for the degree
Doctor of Philosophy
in
Computer Science and Engineering

May 2000

The dissertation "Automatic Time Alignment of Phonemes Using Acoustic-Phonetic Information" by John-Paul Hosom has been examined and approved by the following Examination Committee:

Dr. Ronald A. Cole
Professor
Thesis Research Adviser

Dr. Yonghong Yan
Associate Professor

Dr. Hynek Hermansky
Professor

Dr. John Lazinchbury
Associate Professor

Dr. Wayne H. Ward
Research Professor
University of Colorado at Boulder

Acknowledgements

This work was made possible by the support of a large number of people (and institutions). I would like to formally thank the following: Ron Cole, who has obviously been a very influential and guiding force in the motivation, support, development, and conclusion of this work; Andrew Cronk, for discussions and for asking good questions; The CSLU Member companies, who support CSLU and our efforts; Jacques de Villiers, for countless assistance and clairvoyant insight in using the CSLU Toolkit, the network, and computers in general; Terry Durham, for manually labeling the OGI Kids' Speech corpus; Mark Fanty, for his advice and help with training neural networks; Hiroya Fujisaki, for his insightful and wide-ranging advice, suggestions, and lessons, and for sending his paper on F0 estimation; Jim Glass, for answering numerous questions about the SUMMIT system; and Peter Heeman, for asking as well as answering questions.

In addition, I would like to thank Hynek Hermansky, for his probing questions and insightful advice; Alexander Kain, for discussions and assistance of all sorts; Ed Kaiser, for discussions about the details as well as the big picture; John Launchbury, for taking the time to review my work as someone outside of the field; Chaojun Liu for help with Mandarin phonetic representations; Todd Leen, for advice on training neural networks and for answering questions; Mike Macon, for asking good questions, answering my own questions, and providing the MWM-EGG corpus and associated pitch marks; Dom Massaro, for answering questions and sending me papers of interest; and The National Science Foundation, for the grants (GER-9354959 and IRI-9614217) that supported this work (although the views expressed in this thesis do not necessarily represent the views of the NSF).

Finally, I would like to thank Mike Noel, for discussions and help in obtaining corpora; Bryan Pellom, for his assistance and advice on forced alignment; Joe Picone, for the

phoneme-level manual transcriptions of the Switchboard corpus; Johan Schalkwyk, for his help with the CSLU Toolkit and signal processing; Kal Shobaki, for countless assistance in using the CSLU Toolkit, the network, corpora, and computers in general; Pieter Vermeulen, for asking and answering questions; Sarel van Vuuren, for the modified HTIMIT corpus; Wayne Ward, for answering questions and providing advice; Johan Wouters, for in-depth discussions and assistance of all sorts; Mikio Yamaguchi, for his advice and guidance; and at the end of this alphabetical list but certainly not least in my gratitude, Yonghong Yan, for his excellent advice and instruction.

If this work has at all succeeded, its success rests on the efforts of these people and groups.

Contents

List of Tables

List of Figures

Abstract

Automatic Time Alignment of Phonemes
Using Acoustic-Phonetic Information
John-Paul Hosom

Ph.D., Oregon Graduate Institute of Science and Technology
May 2000

Thesis Advisor: Dr. Ronald A. Cole

One requirement for researching and building spoken language systems is the availability of speech data that have been labeled and time-aligned at the phonetic level. Although manual phonetic alignment is considered more accurate than automatic methods, it is too time consuming to be commonly used for aligning large corpora. One reason for the greater accuracy of human labeling is that humans are better able to locate distinct events in the speech signal that correspond to specific phonetic characteristics. The development of the proposed method was motivated by the belief that if an automatic alignment method were to use such acoustic-phonetic information, its accuracy would become closer to that of human performance. Our hypothesis is that the integration of acoustic-phonetic information into a state-of-the-art automatic phonetic alignment system will significantly improve its accuracy and robustness.

In developing an alignment system that uses acoustic-phonetic information, we use a measure of intensity discrimination in detecting voicing, glottalization, and burst-related impulses. We propose and implement a method of voicing determination that has average accuracy of 97.25% (which is an average 58% reduction in error over a baseline

system), a fundamental-frequency extraction method with average absolute error of 3.12 Hz (representing a 45% reduction in error), and a method for detecting burst-related impulses with accuracy of 86.8% on the TIMIT corpus (which is a 45% reduction in error compared to reported results). In addition to these features, we propose a means of using acoustics-dependent transition information in the HMM framework. One aspect of successful implementation of this method is the use of distinctive phonetic features.

To evaluate the proposed and baseline phonetic alignment systems, we measure agreement with manual alignments and robustness. On the TIMIT corpus, the proposed method has 92.57% agreement within 20 msec. The average agreement of the proposed method represents a 28% reduction in error over our state-of-the-art baseline system. In measuring robustness, the proposed method has 14% less standard deviation when evaluated on 12 versions of the TIMIT corpus.

Chapter 1

Introduction

1.1 Motivation

A fundamental assumption in much of speech processing is that the basic unit of speech is the phoneme. Most speech recognizers identify words based on their phonetic representation, and nearly all speech synthesizers concatenate or synthesize waveform segments according to phonetic pronunciations. In addition, synthetic facial animation of words is usually done at the level of "visible phonemes," or *visemes*, which are closely related to phonemes. As a result, one requirement for researching and building spoken language systems is the availability of speech data that have been labeled and time-aligned at the phonetic level. In addition, time-aligned labels can be useful in language training, analysis of hearing disorders in children, and studies of coarticulation and prosody.

Time-aligned phonetic labels can be created either by a trained human labeler or by an automatic method. Although precise evaluation of the quality of phonetic labeling is difficult, there is a general consensus that manual labeling is more accurate than automatic labeling; this consensus can be seen in the following statements by researchers in the field: Andrej Ljolje notes that "due to the ... inherent limits in the parameterization of the speech signal and the speech model structure, the accuracy of the transcription [by automatic methods] is inferior to that achieved by human transcribers" [92]. Piero Cosi states that "The accuracy of automatic alignment systems will always be checked using references manually segmented by phonetic or speech communication experts" [32]. Stephen Cox reports that "It is well known that ... variation [of manual alignments] is generally small when compared with alignments produced by automatic systems" [33].

To give further weight to the claim that manual alignments are more accurate than automatic alignments, systems that depend on alignment information can be developed using both methods, and the performance of the two systems can be compared. In one case, a speech synthesizer was created using both manually-aligned and automatically-aligned labels; the speech quality of the manually-aligned system was judged in listening tests to be superior to the quality of the automatically-aligned system [33]. In another case, a speech recognizer trained using manually-aligned phonemes had an 11% reduction in word-level error and a 12% reduction in sentence-level error compared to an identical system that was trained using automatically-aligned phonemes [66]. (This result is statistically significant, with p=0.002.)

Although manual alignment is considered more accurate than automatic methods, it is too time consuming and expensive to be commonly used for aligning large corpora. Manual segmentation has been reported to take between 11 and 30 seconds per phoneme [83, 89], whereas automatic segmentation can require between 0.1 and 0.4 seconds per phoneme on a Pentium Pro 200 MHz computer. This difference of several orders of magnitude can only become greater with even faster computer performance and better algorithms, while human performance will likely remain the same. In addition to the greater time required to generate manual alignments, there is variability in manually-generated alignments due to the subjective judgement of the human labeler [136, 139, 16, 114, 92]. Because of these disadvantages to manual alignment, "there is a need for a fast, inexpensive, and accurate means of obtaining time-aligned phonetic labeling of arbitrary speech" [149].

The topic of this thesis, then, is the development of a method of performing phonetic alignment that is significantly more accurate and robust than current automatic methods and significantly faster than manual alignment. The principles used to develop such a method may then be applied to other aspects of speech processing, such as speech recognition, speech synthesis, or facial animation.

1.2 General Overview of Current Methods

As noted above, the most accurate method of creating time-aligned phonetic labels is to employ a trained human labeler. This person typically generates phonetic alignments using a software tool that displays the speech waveform, spectrogram, label, and possibly other information. The labeler aligns the phonetic labels with the speech by listening to segments of the waveform and by using knowledge of the relationship between the waveform, its spectrogram, and its phonetic content. As a result, training in phonetics and spectrogram reading is required to produce acceptable label alignments, and manual alignment is a resource-intensive method.

The most common automatic method for aligning speech is called "forced alignment." In this method, recognition of the speech signal is performed with the search result constrained to the known sequence of phonemes. Because the search procedure yields the locations of the phonemes as well as their identities, the phonetic alignment can be obtained by constraining the search in this way. These systems are called "forced alignment" systems because the alignment is obtained by forcing the recognition result to be the proposed phonetic sequence; this phonetic sequence is determined in advance by a pronunciation dictionary, grapheme to phoneme rules, or a human. In general, there is a strong link between automatic speech recognition and forced-alignment techniques, in that the same general processes can often be used for both tasks. Forced alignment and other methods of alignment will be covered in more detail in Chapter 3.

Figure 1.1 illustrates how a state-of-the-art forced-alignment method compares with manual alignment on the utterance "I mean that's abs[olutely unprecedented]" from cellular-telephone speech. Based on visual inspection of the acoustic-phonetic cues in the spectrogram and by listening to each labeled segment, it is clear that the manual alignments are better than the automatic alignments at the boundaries of the /m/, /θ/, /t^h/, and /b/. (Note that International Phonetic Alphabet (IPA) symbols are used throughout the text of this thesis to represent phonemes, and Worldbet phonetic symbols [64] are used in the figures; a list of IPA and Worldbet symbols and example words is given in Appendix C). However, automatic alignment of this speech segment took less than 3

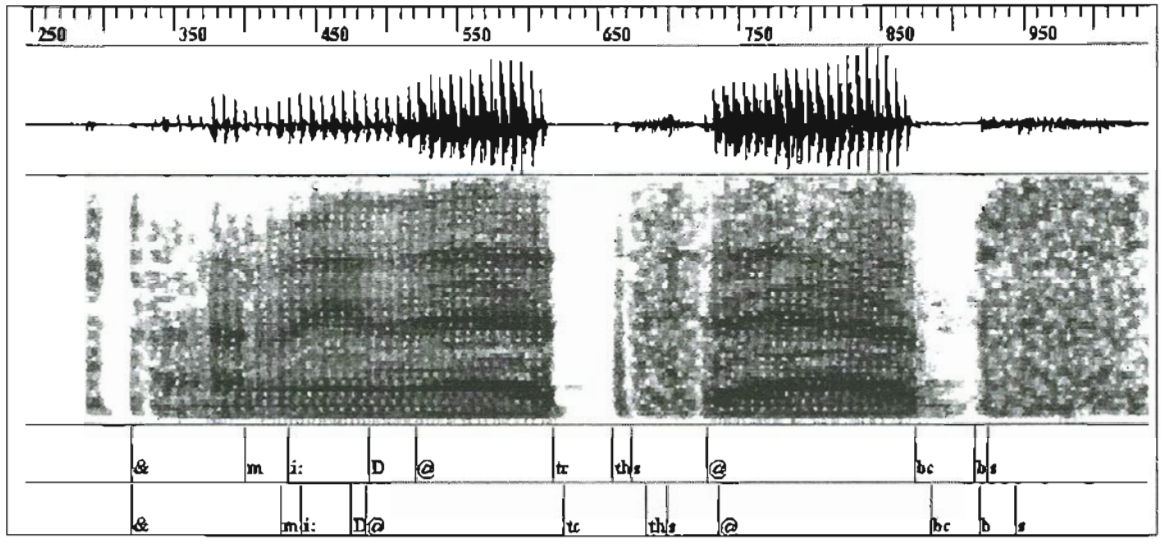

Figure 1.1: Illustration of manual alignment compared with state-of-the-art automatic alignment on the utterance "I mean that's abs[olutely unprecedented.]" Each panel shows, from top to bottom: (a) time marks, (b) the waveform, (c) the spectrogram, (d) manual alignments, and (e)results of automatic alignment.

seconds of computer time, whereas manual alignment would take nearly $2\frac{1}{2}$ minutes of human effort at the reported rate of 11 seconds per phoneme.

1.3 Overview of Proposed Method

A system that does ideal phonetic alignment would have the following attributes:

- accuracy of human labelers at identifying important phonetic events and at working under various channel, noise, and speaker conditions,
- the internal consistency of an automatic method, and
- operation in real time or faster.

This thesis presents the results of an initial effort at building a system that meets these requirements, the methods used to obtain these results, and the motivations for the given

methods.

One reason for the greater accuracy of human labeling over automatic methods is that humans are better able to locate distinct events in the speech signal that correspond to specific phonetic characteristics, such as the sudden increase in energy that signals the beginning of a plosive. This acoustic-phonetic information may provide robustness under conditions of channel distortion, speaker variability, and noise. The development of the proposed method was motivated by the belief that if an automatic alignment method were to use such acoustic-phonetic information, its accuracy would become closer to that of human performance, while still maintaining the internal consistency of current automatic methods. Our hypothesis is that the integration of acoustic-phonetic information into state-of-the-art automatic alignment systems will significantly improve the accuracy and robustness of phonetic alignment.

The model for the proposed method uses standard forced alignment as a foundation. This model then incorporates specific acoustic-phonetic features into the stochastic phoneme-classification process, performs alignment based on classification of phonetic transition regions as well as classification of context-dependent phonemes, and uses phonetic theory to address the issue of data sparsity. Previous work on integrating acoustic-phonetic information into the speech recognition process has indicated that current methods of automatically extracting such information may not yield results that dramatically improve recognition performance. For example, in Schmid's work [130], the use of computed formant trajectories, formant amplitude, formant bandwidth, fundamental frequency, segment durations, and standard cepstral-domain features as input to a phonetic classifier resulted in an 8.8% reduction in error over the use of standard cepstral features alone. If the correct formant values were used instead of the estimated formant values, then the reduction in error increased to 17%. In work by Hosom prior to this thesis [65], voicing information was used as an additional feature for the digits classification task. The use of voicing values extracted by a voicing-estimation algorithm resulted in an 9.8% reduction in error, which is about half of the 19% reduction in error that could have been obtained if the correct voicing values were used. These results indicate that a significant issue in implementing a feature-based system is accurate extraction of the features that

are used. As a result, new methods of extracting acoustic-phonetic features have been developed as part of this thesis.

1.4 Evaluation Methodology

One issue in developing an automatic alignment system is the metric by which success is measured. There is no known method to assess the correctness of a given phonetic alignment, because the exact times at which phonemes begin and end can sometimes be a subjective decision. Not only will different human labelers disagree about the exact placement of a boundary, but a single human shows variability in boundary placement. Because neither human nor machine alignments can be considered completely accurate, it is not possible to compute an absolute measure of accuracy.

The most common method of measuring the performance of an automatic alignment system is to assume that manually-generated labels are correct, and to compute the automatic-alignment error relative to these values [3, 9, 15, 16, 19, 35, 36, 89, 93, 92, 96, 113, 115, 114, 123, 134, 137, 136, 139, 142, 147, 149, 141]. This is an acceptable method if the difference between the automatic alignments and the unknown correct alignments can be considered greater than the difference between the manual alignments and the unknown correct alignments, and if we accept that an automatic system that produced no actual errors might still have a positive error measurement due to the variability in the manual alignments. Because perfect agreement between an automatic alignment and a manual alignment is unrealistic (and in such a case the automatic system would have the same internal variability as the human), performance measured in this way can not be considered accuracy, but simply agreement. Although the term "accuracy" is commonly used in the literature, we will use the term "agreement" to signify the distinction.

A second method of measuring performance determines the robustness of the automatic alignment system instead of its agreement with manual alignments. In this method, we use a test corpus that has been subjected to a series of channel distortions. If we automatically align the speech under these different channel conditions, we can measure the variability in alignment performance without making reference to manual alignments. Because the

true phonetic boundaries do not change with different channel conditions, any changes in the automatic alignment values indicate a lack of robustness. The amount of change under different conditions is considered to be inversely correlated with the degree of robustness of the automatic method.

For this thesis, we will compare the agreement and robustness of our proposed system with a baseline forced-alignment system. Success will be based on whether or not the proposed system has significantly better agreement and is significantly more robust.

1.5 Summary of Research Issues

In summary, the problem addressed in this work is that of aligning the phonetic content of speech with its corresponding acoustic signal. The hypothesis is that specific acoustic-phonetic information can be used by an automatic alignment system to significantly improve performance compared to a baseline automatic system, where the measures of success are agreement with human performance and robustness. Our approach is to measure specific acoustic-phonetic information at different levels of the forced-alignment process, and integrate this information in a probabilistic manner into the standard stochastic framework.

1.6 Outline

The following topics will be addressed in this thesis:

- Models of speech,
- Previous work in automatic alignment,
- Description of the baseline system,
- Overview of the proposed approach,
- Acoustic-phonetic features,
- Integrating transition information,

- Using distinctive features,
- Evaluation methodology, and
- Results and Discussion.

In addition, Appendix A provides a description of the stochastic frame-based recognition systems (HMMs and HMM/ANNs) that are the foundation for many alignment systems, Appendix B contains a glossary of speech-related terms, and Appendix C contains a list of phonetic symbols used throughout this thesis.

Chapter 2

Models of Speech

In order to develop a phonetic alignment method that incorporates acoustic-phonetic information, it is important to understand theories of human speech production and perception, as well as current approaches to computer speech recognition. This section provides background on some of the more prevalent models of speech production and recognition, and in doing so also provides a foundation for concepts and terminology in speech processing.

2.1 Models of Speech Production

2.1.1 The Source-Filter Model

The most common model of speech production is called the "source-filter" model, which was proposed by Johannes Müller in 1848 and described in detail by Gunnar Fant [46]. In this model, the production of speech is composed of three independent parts: a sound source (often the vibration of the vocal folds), a tube through which this sound source passes (usually the vocal tract), and radiation of the sound from the mouth. This model can be used to describe the speech signal in the spectral domain, at one instant in time.

There are several possible sound sources in speech production. The most common is vibration of the vocal folds, which occurs when we utter vowels, nasals, retroflex sounds, liquids, and glides. These sounds, such as /ɛ/ and /m/, are called voiced sounds. The vibration of the vocal folds creates a series of energy pulses; this pulse train has a spectral slope of approximately –12 dB/ocatave. Several models of this source of voiced speech have been developed (for example [48, 47, 81]). In general, these models describe an increase in air flow as the glottis opens, a more sudden decrease in air flow as the glottis

closes, and no air flow while the glottis remains closed for the remainder of the pulse cycle.

A second sound source is frication, which is produced by forcing air through a narrow constriction in the mouth. Frication is the sound source for phonemes such as /f/ and /s/. Fant states that this source has a spectral slope of –6 dB/octave [46], and frication is often modeled using a random-noise generator.

A third source of sound is plosion, which is produced by building up air pressure behind an obstruction (such as the tongue or lips) and then quickly removing that obstruction; this results in a sudden burst of air being released from the mouth. Phonemes that include plosion as a sound source are the plosives and affricates, such as /p^h/, /t^h/, /d/, and /ʧ/. The source for plosion is usually modeled as a sudden step-like increase in air pressure with subsequent slow decay, resulting in a nearly flat spectral shape.

The phonetic identity of a sound is further developed as the sound source passes through the vocal tract or nasal cavity. Each phoneme is produced with a certain position of the tongue and jaw, and these positions determine the shape of the vocal tract. Different shapes of the vocal tract have different resonant frequencies, and these resonant frequencies are called formants. Given a particular vocal tract length and shape, the formant values can be computed, and these formant values can then be represented by filter parameters. In this way, the sound source is filtered as it passes through the vocal tract, where the frequencies that are emphasized are dependent on the phoneme being produced.

In the final stage of the source-filter model, the speech sound is radiated from the mouth. The effect on the spectrum caused by the radiation of sound is called the *radiation characteristic.* The shape of the mouth opening is approximated by a point source, which results in the spectral slope of the speech being increased by +6 dB/octave. In some implementations of the source-filter model (such as formant-based text-to-speech synthesis), the sound source and radiation characteristic are combined into a single representation. This results in a voiced-source spectral slope of –6 dB/octave, and flat spectral slopes for the fricative and plosive sources.

The source-filter model is quite powerful in describing several characteristics of speech, such as the overall spectral shape of sounds and the locations of formants based on the shape of the vocal tract. This model is also successful in explaining effects such as the

overall increase in formant frequencies for female speakers, which is due to the difference between the typical male and typical female vocal tract lengths.

2.1.2 Time-Based Modeling of Speech

The source-filter model describes the speech signal at one instant in time; speech, however, is characterized by change over time in the sound source and resonant frequencies. Several properties and models of the time-dependent nature of speech are outlined here.

At the physical level, the rate of speech is governed by the inertia of the articulators. The body of the tongue moves relatively slowly, and the rate of sonorant phonemes is limited by the rate at which the tongue moves. The lips and tip of the tongue can move faster, and so plosive sounds occur over a much shorter time interval. At the phonetic level, the average duration of phonemes ranges from 20 msec for voiced plosives (/b/, /d/, /g/) to 150 msec for diphthongs (/aɪ/, /ei/, /iʊ/, etc.), with an average phonetic duration of about 70 msec. In addition to durational variation due to phonetic differences, vowel duration may change by a factor of eight, depending on speaking rate, syntax, and stress [79]. Kanedera and Hermansky [72] have studied the perceptually-important modulation frequencies in speech, and found that most of the important temporal change in the speech signal occurs at 4 or 5 Hz, or about every 200 to 250 msec, which is approximately the duration of one syllable [59]. Finally, in recognition of speech, duration information is used by humans to distinguish long from short vowels, voiced from unvoiced fricatives and consonants, phrase-final from non-phrase-final syllables, and stressed from unstressed vowels [79].

The many factors that influence speech duration and the many uses of duration in human speech perception result in fairly complex models. In one model proposed by Klatt for speech synthesis, seven factors that influence the durational structure of a sentence are specified, and there are eight rules that account for these seven factors. This model is, as noted by its author, "only a preliminary step toward a complete theory" [79]. A simpler model proposed by van Santen [140] is able to account for 86% of the variance of vowel durations in a large corpus of manually-segmented speech. This model requires eight parameters, controlling the factors of intrinsic vowel duration, pitch accent, syllabic

stress, post-vocalic consonant, pre-vocalic consonant, within-word position, and utterance position. In a statistical-based model for speech recognition [22], a multi-level sub-lexical tree (called the ANGIE framework) is used to model duration from the phone level up to the word level. A data-driven approach has been used to estimate duration factors at each sub-lexical node in the tree. The duration information contained in the tree can then be used to test various word hypotheses, and favor those hypotheses that have a better match to the model durations. Using this model, an 8% reduction in error on a continuous-speech recognition task was obtained, with a 22% reduction in error for a word-spotting task.

A second characteristic of speech as it changes over time is called coarticulation. Coarticulation is the effect that one phoneme has on its neighboring phonemes; this effect is manifested as a smooth change in formant frequencies from one phoneme to the next. This smooth transition between phonemes is one of the main factors that makes it difficult to determine the exact location of a phonetic boundary.

Several models of coarticulation have been proposed. In a model developed by Öhman [112], coarticulation in vowel-consonant-vowel (VCV) utterances is expressed in terms of vocal-tract shape by the formula

$$s(x,t) = v(x) + k(t)[c(x) - v(x)]w_c(x) \tag{2.1}$$

where $s(x,t)$ is the shape of the vocal tract at a position x and time t, $v(x)$ is the vocal-tract shape corresponding to a given vowel, $c(x)$ is the vocal-tract shape of the consonant, $k(t)$ is an interpolation term that ranges from 0 to 1, and $w_c(x)$ is a term that describes the extent to which $c(x)$ "resists" coarticulation. This model is successful in describing context-dependent variations in vocal-tract shapes using context-independent descriptions of the vowels and consonant. However, this model was only evaluated on VCV utterances, and Öhman briefly noted the seven modifications that would be necessary to model coarticulation of general speech. Öhman also noted the difficulty in his model for describing coarticulation between consonants, such as in a consonant-vowel-consonant (CVC) utterances.

In the "locus theory" of coarticulation [40], consonants are assigned fixed formant values that may not be visible in the speech signal; these "virtual" formant values are

interpolated with the formants that appear in vowels to create the context-dependent formant changes seen in speech. Klatt modified the locus theory so that the interpolation depends on the type of vowel [80]. Using this method, he achieved a consonantal intelligibility of 95% for synthetic CVC syllables, as compared to the intelligibility of 99% for natural-speech CVC syllables. Klatt did not, however, evaluate this model on continuous speech, in which coarticulation effects may extend over a duration of up to six phonemes [74].

In the model proposed by Löfqvist, as reported in Cohen and Massaro [23], speech segments have overlapping "dominance functions" that control the articulators, with one dominance function per articulator. The dominance functions can differ in time offset, duration, and magnitude, giving relatively more or less weight to articulators associated with a given speech segment. Although this model is quite successful at modeling visual speech (in which the articulators are direct parameters of the system), it is not obvious how this model could be used directly in current speech recognition systems, in which the articulators are at best indirect parameters.

A review of six theoretical models that describe coarticulation in continuous speech was conducted by Kent and Minifie [74]; their conclusion was that "coarticulatory patterns are not explained adequately by any of the theories or models discussed herein." This conclusion highlights the complex nature of coarticulation and the difficulty of developing accurate models. Also, in considering these models for use in speech recognition or alignment, it is important to note that even in the simple case of CVC or VCV utterances, it is not possible to easily reverse Öhman's, Klatt's, or Löfqvist's equations to derive context-independent representations from the context-dependent acoustics.

2.2 Models of Human Speech Recognition

There are several models that describe human speech recognition at various levels of detail and at different levels of the speech recognition process. In this section we briefly describe some of the more prevalent models.

2.2.1 The Motor Theory of Speech Perception

The Motor Theory of Speech Perception (abbreviated as the "motor theory") is one of the most widely-cited theories of human speech perception. This theory states, in a more recent version [90], that "the objects of speech perception are the intended phonetic gestures of the speaker, represented in the brain as invariant motor commands that call for movements of the articulators." According to this theory, when we perceive speech, we perceive the gestures that correspond to the articulatory movements of the speaker, such as lip rounding and jaw raising. Furthermore, in this theory there is a "specialized module" in the brain that translates from the acoustic signal to the intended articulatory gestures. According to Liberman, such a module might work using the analysis-by-synthesis method [6], in which a mental model of a speech synthesizer is used to generate various acoustic properties. The acoustic-gesture parameters that are input to this synthesizer are varied until the error between the synthesized acoustic properties and the observed acoustic properties is minimized. The resulting articulatory gestures are the output of this module. Liberman and Mattingly claim that "the processes of speech perception are ... inherently computational and quite indirect. If perception seems nonetheless immediate, [it is because] the module is so well-adapted to its complex task."

There are several criticisms of the motor theory concerning a number of its aspects. Cole et al. [29] have refuted the claim that there is a "biologically based link between perception and production ... [that] occurs only in speech" [90]. Cole showed that the use of printed spectrogram displays can be interpreted by the eye and used to classify the phonemes in continuous speech with at least 85% accuracy; single-word utterances can have a phonetic classification rate of at least 93%. Such visual reading of speech without a biological "specialized module" argues against the necessity of such a module when aurally recognizing speech. Furthermore, the person who read the spectrograms in that study did so without making explicit reference to articulatory gestures. This work challenges the claims that the acoustic signal is too complex to be directly mapped to phonetic categories, and that human speech perception requires the intermediate stage of determining articulatory gestures.

In another criticism of the motor theory, Lane investigated the use of CV speech stimuli that had been modified so that the formant frequencies were inverted on the frequency axis. Such stimuli are heard as non-speech sounds, even though they have the same temporal patterns found in ordinary CV stimuli. Lane trained subjects to classify these modified stimuli, and found that "the categorization ... of speech cues [is] not necessarily due to the operation of a special motor reference process, because the same results can be obtained, after proper auditory training, for stimulus differences that are not producible by speaking" [117].

Finally, as Ladefoged points out, we are able to perceive two speech sounds as being the same, even if the articulator positions and movements used to produce the sounds are different. This occurs in "r-colored vowels," which can be produced with the tip of the tongue up or with a raised tongue position further back in the mouth. This effect can also be obtained in the production of rounded vowels, such as /u/, with a lowered larynx position or with increased lip rounding. In both of these cases, although the articulators are in different positions, the resulting acoustic properties are the same. The fact that these sounds are perceived to be the same can be more parsimoniously explained by an auditory-based theory of speech perception than by the theory that listeners perceive speech "by reference to their own motor activities" [84].

2.2.2 The Multiple-Cue Model of Speech Perception

In contrast to the motor theory, Ronald Cole and Brian Scott proposed a model of speech perception in which a combination of context-independent invariant cues and context-dependent phonetic transition cues are integrated when recognizing syllable units [28]. Properties of the waveform envelope are also used when integrating syllables into higher-level units such as words and phrases. This model will be referred to here as the "multiple-cue" model.

Cole and Scott provided evidence for invariant cues in all consonant phonemes. These invariant cues may uniquely identify the phoneme (as in the case of /s/, /z/, /ʒ/, /ʃ/, /ʧ/, and /ʤ/), or they may be used in conjunction with transition cues to identify the phoneme (as in the case of /f/, /θ/, /v/, /ð/, /m/, /n/, and /ŋ/). In the case of stops,

the voiced/unvoiced distinction (/b/, /d/, /g/ as opposed to /p^h/, /t^h/, /k^h/) is signaled by invariant cues, while the place of articulation involves either invariant or transitional cues. In addition to these two types of cues, properties of the waveform envelope can be used to recognize prosodic as well as phonetic information.

This multiple-cue model is attractive in that (a) it is computationally more feasible than the inherently complex motor theory, (b) there is a direct mapping from acoustics to phonetics, making the speech signal more amenable to analysis, and (c) it accounts for aspects of the waveform that are both invariant and context-dependent. In criticism of this model, the time-domain waveform amplitude does not seem to be a likely candidate for human speech recognition, as the waveform signal is directly converted into a frequency-domain representation by the cochlea. However, the information that Cole and Scott determined using the time-domain waveform (amplitude, pitch, and duration) may also be extracted from a time-varying spectral representation.

2.2.3 Invariant Cues for Stop Perception

A study of speech perception that focused on one aspect of human performance was conducted by Stevens and Blumstein [133]. The result of this research identified an acoustic-phonetic cue that can be used to uniquely identify the place of articulation in stop consonants, based on human perception of synthetically-generated consonant-vowel phonemes. This cue is the gross spectral shape of the consonant, sampled at both the burst onset and the voicing onset. This work gives support to Cole and Scott's multiple-cue model, in specifying an invariant cue that can be used for identifying place of articulation in stops.

2.2.4 The Fletcher-Allen Model

Between 1918 and 1950, Harvey Fletcher and his colleagues studied human speech perception at Bell Labs. As a result of this effort, they developed a theory of human speech recognition that has been elaborated upon more recently by Jont Allen [1]; we will summarize a few of their contributions here. One result of Fletcher's work was measurement of correct CVC syllable recognition in terms of recognition rates of the component phonemes:

$$S = c_1 v c_2 \tag{2.2}$$

where S is the probability of correct identification of the CVC syllable, c_1 is the probability of correct recognition of the first consonant, v is the probability of correct recognition of the vowel, and c_2 is the probability of correct recognition of the second consonant. This formula has the important implication that humans perceive each phoneme individually, rather than the syllable as an entire unit. In addition, Fletcher found evidence that humans process frequency bands independently, and that the overall error for recognition of several bands is equal to the multiplication of errors in each individual band.

Allen interprets this to mean that humans perform partial recognition of frequency bands independently, and that these partial results are "fused" to produce estimates of phonemes. In general, the number of frequency bands should be between 10 and 30. Allen also notes that an important transformation takes place within each band, namely that "the neural representation of the signal intensity has been transformed into a measure of partial recognition ... we must not assume that this is a trivial transformation" [1].

Based on Fletcher's findings, Allen proposes a cascaded model of human speech perception, in which the acoustic signal is first broken into heavily-overlapped frequency bands. The outputs of these bands are used to extract "phone features" in about 20 different bands, which are then used to classify phones. The phone-level classification is then used to classify syllables, which are in turn used to classify words. Allen also notes that "[it is] unlikely that feedback is common or significant between the deeper layers and the outer layers" [1].

2.2.5 Auditory Scene Analysis

Auditory Scene Analysis (ASA) [13] is a theoretical model of human speech perception in which both bottom-up and top-down processing are used to determine what parts of the speech signal belong to a single acoustic event. As a result, ASA tends to focus on complex auditory environments involving multiple sounds. Often, grouping of patterns in the speech signal into "streams" is done on the basis of similarity, in pitch [30] or other aspects. This model is able to explain why a signal may be interrupted by a brief, stronger signal, but still be perceived as continuous. In a computational model of ASA, Cooke and Brown are able to detect certain occluded sounds and restore them (their example being

speech occluded by a siren) [30].

2.2.6 The TRACE Model

The TRACE model of speech perception was developed by James McClelland and Jeffrey Elman in 1986 [101]. This model was designed to be implemented on a computer while still being a plausible model for human speech recognition. The TRACE model has three levels: the feature level, the phoneme level, and the word level. The feature level is composed of seven distinctive features (consonantal, vocalic, diffuseness, acuteness, voicing, power, and amplitude of burst noise), each of which can have one of nine values. Each level is constructed by connecting a number of simple processing units, and recognition "takes place through the excitatory and inhibitory interactions of a large number of [these] units, each working continuously to update its own activation on the basis of activations of other units to which it is connected." Each unit represents a hypothesis about the input, with the activation of the unit monotonically related to the strength of the hypothesis; the connections between units represent relationships between hypotheses. Units on the same level that are inconsistent have mutually inhibitory connections. The connections between layers are bi-directional, which allows both bottom-up and top-down processing to occur simultaneously.

The TRACE model is able to account for a number of effects observed in human speech perception, giving support to the psychological validity of this model. These effects include top-down lexical effects (a faster response to words than to non-words), the perception of phonemes as distinct categories instead of having continuous values, and results consistent with phonotactic rules (such as /sl/ being a valid phonetic combination, whereas /sɻ/ is not), even though such rules were not explicitly provided. McClelland and Elman list eleven similarities between TRACE and human speech recognition, but also note that "although TRACE has had a number of important successes, it also has a number of equally important deficiencies," most of which are related to simplifying assumptions in the implementation of the model.

The TRACE model has obvious parallels to artificial neural networks, with one important difference being that in the TRACE model, the connections are bi-directional,

whereas in typical feed-forward or recurrent networks, the connections are uni-directional. In addition, McClelland and Elman had no formal means of training the system, and relied on hand tuning of the parameters to obtain their results.

2.2.7 The Fuzzy-Logic Model of Perception

The Fuzzy-Logic Model of Perception (FLMP) is not a complete model of human speech recognition, in that it does not specify all of the steps from input of the speech signal to output of the recognized words. This model focuses on the integration of feature information to arrive at classification results that are consistent with human performance.

The FLMP consists of three stages: feature evaluation, feature integration, and pattern classification. In the feature evaluation stage, the speech signal is analyzed, and certain features are extracted. For example, there may be a feature called "labial" to indicate a place of articulation. The values of these features are continuous, and they represent the degree of belief that the current speech segment indicates the specified feature. For example, the value for the "labial" feature may be 0.80, indicating a reasonably strong belief that the speech segment consists of a labial sound. The use of continuous values for each feature is supported by various studies of human speech perception [111].

The second stage consists of prototype matching, in which the input is matched to a prototype description of each possible phoneme. For example, the phoneme /b/ may have the prototype features "labial" and "voiced." The phoneme prototypes are specified by "matching functions" in terms of fuzzy-truth values, so that, for example, the matching function for /b/ may be specified as

$$B_s = L_s \, V_s \tag{2.3}$$

where B_s is the degree to which the perceived speech, s, will match the phoneme prototype for /b/, L_s is the degree of belief that the speech is labial, and V_s is the degree of belief that the speech is voiced. (In an extended version of the FLMP, the belief values are modified by exponential weights that indicate the importance of extreme values of that feature.) The extent to which the input speech matches each prototype is computed by evaluating all of the matching functions.

In the third stage, pattern classification is performed. The probability of identification of each phoneme is computed using Luce's model [95]. For our example of /b/, the probability of the speech containing a /b/, if the only possibilities are /p^h/, /b/, and /d/, is given by:

$$p(/b/|s) = \frac{B_s}{(P_s + B_s + D_s)} \tag{2.4}$$

where $p(/b/|s)$ is the probability of a /b/ given the speech signal s, P_s is the matching function for /p^h/, B_s is the matching function for /b/, and D_s is the matching function for /d/.

Massaro and Friedman used the FLMP to model the results of several human pattern classification tasks, and found that the FLMP provides an equivalent or better representation of the data when compared with a number of other information-integration models, including additive, linear integration, two-layer connectionist, theory of signal detectability, and multidimensional scaling models [100]. It should also be noted that the FLMP is mathematically equivalent to Bayesian integration if the fuzzy-truth values are interpreted as probabilities, although the FLMP was developed based on psychological studies and without reference to Bayes' rule.

2.3 Models of Computer Speech Recognition

There are a number of models of computer speech recognition, each with a different perspective. Most models can be generally classified as either segment-based or frame-based. In this part, we describe some influential segment-based and frame-based systems.

2.3.1 Segment-Based Systems

The SUMMIT System

The SUMMIT system was originally developed by Victor Zue and his colleagues at MIT in the 1980s, and several variations have evolved over the years under the guidance of Jim Glass. One defining characteristic of the SUMMIT system is that it first divides the speech signal into segments, and then phonetically classifies each segment. The classification

scores of the phonetic segments are searched to find the most likely word sequence. The general procedure for recognition in the SUMMIT framework is as follows:

1. Acoustic boundaries (landmarks) are determined based on the amount of local spectral change. In one extreme implementation of SUMMIT [21], boundaries are placed automatically at every 10-msec frame (effectively transforming SUMMIT from a segment-based to a frame-based system), but this approach is not currently used because of the large amount of required computation time.

2. A network of segments (dendrogram) is created by one of the following methods:

   (a) Merging short segments into longer segments according to their spectral similarity. This is the "traditional" approach used in SUMMIT [57], which has modest computational requirements.

   (b) Segmentation by recognition, in which a recognizer is used to classify each frame or acoustic landmark as a phoneme or a phonetic transition. After this classification, a forward-pass Viterbi search is done, which is followed by a backward-pass A* search. The A* search yields a number of alternative phonetic segmentations of the speech signal; these segmentations form the resulting dendrogram. This approach is computationally more expensive, but yields better recognition performance than the traditional approach. The most recent implementation performs the segmentation in real-time on a 200 MHz CPU [21, 88].

3. Given the dendrogram created in Step 2, phonetic classification of all segments is performed using one or both of the following methods:

   (a) The first method performs context-independent recognition of each segment in the dendrogram. In this method, there are between $N+1$ and $2N$ recognition categories, where N categories correspond to the N possible phonemes, and the remaining categories are used to model segments not included in a hypothesized segmentation (called "not modeling" or "near-miss modeling") [21].

(b) The second method performs context-dependent recognition of each segment boundary in the dendrogram [56]. The context-dependent categories can be phonetic boundaries or phoneme-internal boundaries, and there can be as many as $(N + N^2)$ recognition categories. In practice, only about 750 categories are used.

These classifiers are trained with the same spectral-domain features that are commonly used in HMM speech-recognition systems (described below), and classification is done using mixtures of Gaussians.

4. Searching is done with a bigram forward-pass Viterbi search and, for N-best hypotheses, an n-gram A* backward-pass search. If both context-independent segment recognition and boundary recognition are done in Step 3, then the final probability of a word sequence is the multiplication of the probabilities of the segment result and the boundary result for that sequence.

Performance of the most recent version of SUMMIT is about 72% accuracy on phoneme classification of the TIMIT database. This phoneme-level result is among the best reported; a standard HMM system is reported to have 69.1% accuracy [87], and a recurrent-neural-network approach yielded 73.4% accuracy [128].

The Feature System

The FEATURE system was developed by Ronald Cole, Richard Stern, and Moshè Lasry at Carnegie-Mellon University in the early 1980's. The motivation for the FEATURE system was to enable automatic speech recognition to perform fine phonetic distinctions, such as between /p^h/ and /b/. At the time that FEATURE was developed, frame-based template-matching systems had recognition rates of only about 60% on the "E-set" of confusable alpha-digits (which consists of the set of letters and digits {B, C, D, E, G, P, T, V, Z, and 3}). The original FEATURE system was designed for speaker-independent recognition of the isolated letters "A" through "Z." The following steps are involved in recognition using FEATURE:

1. *Signal Processing*: Given an utterance corresponding to a single letter, signal processing routines are used to extract general information about the signal, such as spectral properties, the fundamental frequency, the number of zero crossings, and energy in various frequency bands.

2. *Segmentation*: Four points are located in the utterance: the beginning of the utterance, the onset of the vowel, the vowel offset, and the ending of the utterance.

3. *Feature Extraction*: About 50 different features are extracted, using the information determined from Steps 1 and 2. These features include the first three formants of the vowel region, the formant trajectories, the maximum and minimum frequencies of each formant, and the duration of aperiodic sound prior to and following voicing. The types of features were selected using visual inspection of different representations of the signal.

4. *Classification*: A decision-tree approach is used to determine the probabilities of each of the twenty-six letters. Each node in the tree represents a group of letters, and lower nodes contain disjoint subsets of higher nodes; the leaf nodes contain the individual letters. At each non-leaf node, the likelihood of the utterance belonging to that node is determined using multivariate Gaussian probability distributions of the feature vectors. Probabilities are computed for all non-leaf nodes in the tree, and the final probability of a given letter is the multiplication of the probabilities of each node leading to that leaf node. Only relevant features are used at each node to reduce the dimensionality of the decision space, and the assumption is made that the sets of features used for classification at each node are independent.

5. *Adaptation*: The Gaussian probability distributions can then be adjusted to better match the expected feature values of an individual speaker. The probability distributions are updated after recognition of each utterance and after receiving feedback from the user as to which letter was actually spoken (supervised adaptation).

Without using speaker adaptation, the FEATURE system has 89% accuracy on isolated letters, and 83% accuracy on the E-set. Compared to frame-based template-matching

systems, the reduction in error on the E-set is greater than 50%. The use of speaker adaptation further reduces the error rate by another 50%, given sufficient data for adaptation.

The FEATURE system was modified and extended by Cole, Fanty, and others to use neural networks for classification and to recognize continuous-speech letters (spoken with or without pauses between words). Results of this new system, called "EAR," are 96% accuracy on microphone speech and 89% accuracy on telephone-band speech [24]. These results represented the state-of-the-art for six years until a sophisticated HMM system achieved 97.3% accuracy on high-quality speech and 91.7% accuracy on telephone-band speech in 1996 [94].

2.3.2 Frame-Based Systems

Hidden Markov Models

By far the most dominant method for automatic speech recognition is currently the *hidden Markov model* (HMM), which has been used for speech recognition since at least 1975 [7]. This method has an elegant mathematical framework that allows data-driven training of speech units such as words, phonemes, or context-dependent phones. The two major reasons for the widespread use of HMMs are that the mathematics are well formulated, moving speech recognition to the well-researched domain of statistical pattern recognition, and that the performance of HMM systems is most often superior to that of knowledge-based approaches.

The details of using HMMs for speech recognition are presented in more detail in Appendix A, but the basic model for HMM-based recognition is that of independent states linked by state transition arcs, as illustrated in Figure 2.1. Each state is associated with a certain linguistic unit, usually a phonetic-based unit, and at any point in time the system is in one and only one state. With each increment in time (usually about 10 msec), a transition is made to another (or the same) state. During recognition, the system estimates the likelihood of being in each state at time t, based on the likelihood of being in each state at time $(t-1)$, the probabilities of transitioning from the previous states to the current state, and the probability of the current state being associated with the speech signal at time t (called an *observation probability*). The likelihood of each word at

time t_w is then the likelihood of being in the final state associated with that word at time t_w; a simple comparison of word-final state likelihoods yields the most likely word given the speech input and HMM configuration.

In general, the speech signal at time t is represented by spectral-domain information over a small (16 msec) time window. Estimation of the observation probabilities is often done using mixtures of Gaussians, although vector quantization (VQ) has also been used. The search through the HMM states to determine the most likely word is done using a dynamic-programming search algorithm called the Viterbi search.

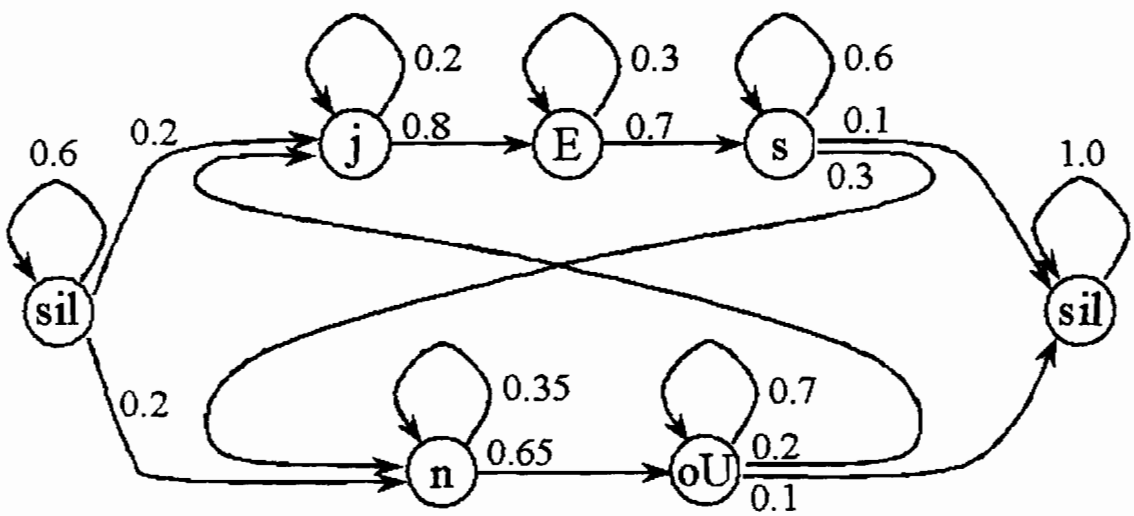

Figure 2.1: HMM state sequence for a two-word vocabulary.

Hybrid HMM/ANN Recognition

Hybrid HMM/ANN systems have been developed at research laboratories such as Cambridge University [128], ICSI [12], and OGI [68]. The main difference between standard HMM-based recognition and hybrid HMM/ANN recognition is in the estimation of the observation probabilities. In standard HMM systems, these probabilities are estimated using mixtures of Gaussians. In HMM/ANN systems, the observation probabilities are

estimated using artificial neural networks (ANNs), which have been shown to estimate *a posteriori* probabilities given sufficient training data and hidden nodes [126]. Other aspects of HMM/ANN systems, such as the state-based framework and the use of the Viterbi search to determine the most likely utterance, are the same. Neural networks have some advantages over Gaussian mixture models (GMMs): ANNs have better discriminative properties, do not require the input features to be uncorrelated, and do not require the data to fit Gaussian models [11].

The main disadvantage of the hybrid HMM/ANN approach is the amount of time required to train a neural-network classifier; this increase in training time is, however, typically offset by a decrease in the time required to estimate the observation probabilities during recognition. Another disadvantage, from an engineering standpoint, is that once the network has been trained, it is not possible to easily modify its properties; with mixtures of Gaussians, phonetic models can be adjusted individually after training. Finally, training the ANN requires good initial estimates of the locations of each phoneme; standard HMM systems can be trained without such information (although even standard HMM performance benefits from the use of manually-aligned transcriptions).

Syllable-Based Recognition

Motivated by psychoacoustic evidence that the syllable is important in human speech perception, Su-Lin Wu, Steven Greenberg, and their colleagues at the International Computer Science Institute (ICSI) developed a syllable-based recognition system [152]. The input to this system is a "modulation spectrogram," which represents the time-varying spectral content of speech with low-pass-filtered independent frequency bands [60]. These bands capture the syllable-length fluctuations in the speech signal while suppressing faster and slower change. The modulation-spectrogram bands are passed to a neural network with a relatively large (185 msec) window of input frames, and are classified into 124 "semi-syllable" categories for the numbers-recognition task. A Viterbi search is then used to search the network outputs for the most likely number sequence.

Performance of this syllable-based system (90.2% word accuracy on telephone-band speech) was not as good when compared with a baseline system (93.2% word accuracy),

but the combined performance of the syllable-based and baseline systems was much greater than either system alone (94.5% accuracy). This indicates that the types of errors made by each system are to some degree independent, and that further research in this area may be fruitful.

Transition-Based Recognition

The Stochastic Perceptual Auditory-event-based Model (SPAM) for speech recognition was developed by Nelson Morgan and his colleagues at ICSI. The motivation for SPAM is that "information needed for correct [phonetic] identification is largely contained in spectral transitions" (called *avents*) [151]. This motivation is supported by the work of Furui, who found that humans associate spectral transitions with phonetic identification in natural speech [51]. In SPAM, the primary unit of classification is the phonetic boundary region, with all non-boundary frames of speech classified as a single "non-transition state." A neural network is used to estimate the probability of a given frame of speech belonging to a particular class, and the most likely word is determined using a Viterbi search.

Recognition experiments using SPAM on the task of digit classification showed that the error rate for SPAM is roughly twice that of standard phonetic recognition on clean speech, and about equal to phonetic recognition performance on noisy speech. However, the combination of results from SPAM and standard phonetic recognition yielded performance much better than either system alone, indicating that the errors in the two systems are independent.

Cravero, Pieraccini, and Raineri [34] created an HMM-based system in which the units for recognition are either context-independent phonemes or diphones, depending on the relevance of the category to recognition (some categories, such as plosives, depend heavily on transition information, and other categories do not). Their general-purpose system has 22 context-independent "stationary" units, and 101 phonetic transition units. On a corpus of nearly 1000 isolated Italian words, an HMM system trained using their proposed categories yielded 77.5% accuracy in word recognition. A comparable system trained on only context-independent phonetic units had 73.7% accuracy in word recognition, and a system trained on only phonetic transitions had 68.2% accuracy. This indicates a clear

advantage to using a combination of phonetic transition and steady-state categories, with a 15% reduction in word-level error for the given task. The authors did not, however, compare the performance of their system with a context-dependent phonetic recognizer.

In related work, Hosom and Cole implemented a transition-based digit recognition system [67], in which the units for classification were phonetic-transition units as well as context-independent phonetic units. As in the SPAM model, classification was done by a neural network, and a Viterbi search was used to determine the most likely word sequence. They found a 13% reduction in error over a context-dependent baseline approach, from 94.7% to 95.4% word accuracy on telephone-band speech. This system was restricted to digit recognition, due to the large number of possible diphones in unconstrained speech.

The SUMMIT system, described above, also performs transition-based recognition during the segmentation-by-recognition procedure, in order to identify the locations of the speech segments [21]. During this procedure, each potential acoustic landmark is classified as either a phonetic transition or a non-transition, and the Viterbi and A* searches are used to determine the N-best segmentations. Transition information can also be used during the classification of segments into phonemes, by multiplying the segmentation likelihoods by the likelihoods of all transitions in that segmentation [56].

Finally, a phonetic alignment system developed by van Santen and Sproat [141] aligns phonemes with speech using phonetic transition information in different frequency bands. This system is discussed in more detail in Section 3.4.

2.4 Human Spectrogram Reading

As noted in the description of the motor theory, humans are capable of recognizing words using only printed spectrograms that display the speech signal along the dimensions of time, frequency, and energy. This ability to recognize speech using spectrogram displays is called "spectrogram reading." (An example spectrogram is given in the third panel in Figure 1.1). The techniques employed in spectrogram reading may be useful in the design of machines that can recognize speech. During spectrogram reading, explicit features in

the speech signal are used, and such explicit features have the potential to be automatically extracted from the speech signal. This potential was a driving force behind the development of the FEATURE system, described above.

The seminal paper on spectrogram reading [29] describes the abilities of a single spectrogram reader, referred to as VZ, and analyzes his approach. It was found that VZ was able to identify more than 97% of all phonetic segments in continuous speech. VZ associated one segment with only one phoneme (one hypothesis per segment) about half of the time, while he associated one segment with two phonemes (two hypotheses per segment) an additional one-third of the time. VZ identified the correct phonetic label 86% of the time. As for his use of higher-level knowledge when reading spectrograms, it was found that labeling performance was actually slightly better on utterances containing nonsense words.

VZ used a two-pass method for reading spectrograms; in the first pass, he segmented the speech, and in the second pass he identified the phonemes in each segment. (There was, however, some adjustment of the segmentation during the second pass.) Segmentation was done primarily by locating the points of spectral change and sharp changes in intensity. Some boundaries were determined based on relative local duration; for example, two adjacent stop closures can be identified as two phonemes even without a change in spectral information, because the combination of both stop closures is significantly longer than the duration of a single stop closure. Changes in formant frequencies (such as a dip in the first formant frequency or a decrease in formant amplitude) were used by VZ to identify transitions within a sonorant region. In some cases, such as liquid-vowel transitions, no boundary was marked until the second pass. VZ used a left-to-right segmentation strategy in some cases, and a less sequential strategy in others. In the less sequential cases, VZ would segment the obvious boundaries first and then identify more difficult boundaries.

Once the speech had been initially segmented, each segment was assigned a phonetic label. Label assignment was done based on knowledge of unique spectral patterns for a phoneme, by knowledge of coarticulatory effects, and by constraints imposed by English phonology. Even for highly complex sounds such as plosives, VZ was able to identify and classify plosives with great accuracy based on the characteristic patterns of the manner

and place of the plosive. Vowel classification was done by first classifying the vowel as reduced or unreduced. For unreduced vowels, the characteristics of the Jakobson, Fant, and Halle set of distinctive phonetic features were used, such as diffuseness and acuteness; these features are related to characteristics such as vowel height or place. Once a general classification had been made, a finer classification was obtained based on relative duration and detection of offglides. It is also interesting to note that by "computing appropriate formant displacements," VZ was able to effectively normalize for the effects of coarticulation. Finally, VZ used phonological rules when necessary. These rules included phonetic-combination rules, such as the lack of the /dl/ combination in English, and rules about allophonic variation that can be predicted based on context, such as /t/ being unaspirated in an /st/ cluster but aspirated (/t^h/) in a word-initial position.

The ability of VZ to read spectrograms and the analysis of his methods has several implications for automatic speech recognition. First, VZ's performance on phonetic classification of continuous speech represents a roughly 50% reduction in error compared to current automatic speech-recognition systems when trained and evaluated on clean microphone speech (with about 85% accuracy for VZ and 70% accuracy for automatic methods). This indicates that the phoneme-level performance of ASR systems can still be greatly improved, which in turn should improve the performance of word-level recognition. Second, VZ did not make use of suprasegmental cues such as pitch contours (for tone and intonation), stress, and rhythm, except for occasional references to local duration and speech rate. If these suprasegmental features were available, the performance of VZ and automatic systems might further improve. Third, for units of classification, VZ "appears to use a mixture of phonemes, diphones, and sub-phonetic units," indicating that the use of all of these units may be advantageous in automatic speech-recognition systems.

2.5 Weaknesses of HMM and HMM/ANN Systems

Although the mathematics that define HMM and HMM/ANN systems are well formulated, the amount of data needed to train an HMM recognizer and the fragility of these systems when used under different conditions indicate that the HMM framework has some

weaknesses when applied to speech recognition. In addition to the well-known issue that independence between the the values of each frame of speech can not be fully justified, there are other ways in which standard HMMs may not be best suited to speech recognition. One major weakness of HMM systems is that because no speech-specific knowledge (other than perceptually-related warping of the spectrogram and data-driven clustering of phonetic categories) is used in computing the likelihoods of each phoneme at each frame, these values are estimates of not only the phonetic qualities of the speech, but also the channel and noise conditions of the training data. (Channel and noise distortions can be addressed by RASTA or CMS pre-processing, but complex distortions, such as are found in telephone speech, are not easily factored out.) In addition, the system is tuned to the characteristics of the speakers in the training set, who may or may not represent the qualities of the speaker(s) in the test set. This union of phonetic and corpus-specific information in training makes recognition performance sensitive to factors that are unrelated to the phonetic content of the signal.

There are several ways in which phoneme-specific knowledge is available but not used in HMM systems. First, one set of features is used to classify all phonemes in an HMM, whereas there is evidence that humans make use of a wide variety of cues. In particular, there is information about voicing, pitch, glottalization, bursts, and intensity that is detectable by humans but not well represented by the standard feature set. As perceptual studies indicate that humans make use of all available relevant cues, it is likely that the use of this information in computing the observation probabilities will improve the robustness of an HMM system.

Second, the observation-probability classifier is trained to classify all frames within a (sub-)phonetic segment as a single (sub-)phonetic category. These estimates will be more unreliable at phonetic boundaries, because the speech signal is changing most rapidly during a transition, there are fewer transition examples than non-transition examples for training, and the data are more widely spread in the feature space due to coarticulation between two phonemes. However, perceptual studies by Furui, the analysis of spectrogram-reading techniques by Cole, and the (sometimes preliminary) speech-recognition and alignment systems developed by Zue, Morgan, Hosom, and van Santen indicate that phonetic

boundaries are well-motivated categories for classification, providing information that is complementary to the phonetic steady-state regions. The use of phonetic transition information in an HMM system may then also lead to more robust recognition, if the problems associated with an inherently large number of phonetic-transition categories can be addressed.

Third, although data-driven clustering of categories into phonetically-related groups is often done in order to make effective use of the training data, the relationships among phonemes that arise out of the physical constraints of the speech production process are not explicitly used. These relationships, if properly encoded, may provide additional structure to an HMM system for improved robustness.

2.6 Summary

There is a wide range of models of speech, depending on the application and the approach of individual researchers. From this array of models, we can draw some general conclusions that may be relevant in the design of automatic speech-recognition and phonetic-alignment systems.

First, although no current automatic system makes use of all relevant information in the speech signal, humans seem to combine information from as many different relevant sources as possible. This belief is seen in Liberman's statement that "every potential cue — that is, each of the many acoustic events peculiar to a linguistically significant gesture — is an actual cue. All possible cues have not been tested, ... but no potential cue has yet been found that could not be shown to be an actual cue" [90]. Or, as Cole and Scott state, "... there are many different cues to each phonetic distinction, and listeners make use of all available cues" [28].

Second, there are several time-domain aspects of speech, notably coarticulation and duration, that are complex and not fully understood. Use of such information may improve automatic speech recognition performance (as was shown in Chung and Seneff's work), but improvements will be limited by the accuracy and robustness of the coarticulation and duration models. Current models, such as those proposed by Klatt, van Santen, Chung

and Seneff, and Öhman, are either still in the research stage or would be difficult to apply to the recognition of speech.

Third, stochastic models of speech recognition usually perform better than rule-based models. This indicates that a small number of rules (or even a larger number of complex rules) does not robustly capture the variability found in continuous speech. However, as the success of statistical models depends on the way in which the data are represented [5], it may be advantageous to utilize some of the known, fixed properties of speech in the design of statistical speech recognizers. This is, however, not easily accomplished; Fred Jelinek, who at the time was with a major IBM speech recognition project, somewhat facetiously claimed that the most effective technique IBM had found for decreasing error rates was to "fire a linguist" [105]. The reason that acoustic-phonetic or linguistic information has not been successfully integrated is not necessarily that our knowledge of linguistics is incorrect, but that it has proven very difficult to extract and reliably incorporate the features that represent this knowledge. As noted by Cole, this difficulty "stands as a major stumbling block to progress" [26].

Fourth, given the various models of human and machine speech recognition, it seems likely that the speech recognition problem is best approached at multiple levels, including specific acoustic features at a lower level, sub-phonetic (possibly including phonetic-transition) units at a higher level, as well as phonetic and syllabic levels. Examples of this hierarchical approach are seen in the TRACE model, as well as in Fletcher and Allen's model. The use of distinctive features in phonetics, in the TRACE model, and in spectrogram reading suggests that distinctive features may also be advantageous in approaches to automatic speech recognition. Results from the use of combined steady-state and transition units by Morgan, Cravero et al., Hosom and Cole, and Glass and Chung indicate the potential of transition information. However, a clear advantage of the transition-based approach over context-dependent, general-purpose HMM or HMM/ANN systems has not been reported in the literature.

Chapter 3

Previous Work in Phonetic Alignment

Of a review of 32 automatic alignment systems, 44% (14 systems) use HMM or HMM/ANN recognition to obtain the alignments, and another 25% (8 systems) use dynamic time warping (DTW). The remaining third (10 systems) employ a wide variety of approaches, including methods that use estimates of voicing [2], measures of spectral variation [116, 139], a hierarchical segmentation structure called a dendrogram [54, 31], diphone detection [141, 73], multiple frequency bands [141], temporal decomposition [144], templates of phoneme sequences [9], and rules that encode acoustic-phonetic knowledge [89].

In this chapter, we will report on the agreement of manual alignments (inter-labeler consistency), discuss the HMM systems, the DTW approaches, a multiple-frequency-band method [141], two diphone-detection methods [141, 73], and a knowledge-based method [89]. Finally, we will describe the current state of the art in automatic phonetic alignment, based on the systems reviewed here.

Automatic-alignment agreement with manual labels is most often reported in terms of what percentage of the automatic-alignment boundaries are within a given time threshold of the manually-aligned boundaries. For example, Brugnara et al. report that for their system, 88.9% of their automatic boundaries are within 20 msec of the manual boundaries. This type of result will be reported here as a percent "agreement" within the given threshold; in this example, Brugnara's system has 88.9% agreement within 20 msec. Results with a threshold of 20 msec will be reported when possible, as this threshold is commonly used and allows general comparison between systems. Relative differences in the agreement between two systems will be reported using the terminology "reduction in error," even if the alternate (although cumbersome) terminology such as "increase in agreement" may

be technically more correct.

3.1 Manual Phonetic Alignment

Evaluation of manual phonetic alignments is subject to the same pitfalls as evaluation of automatic systems. As a result, manual alignment agreement is usually reported as inter-labeler agreement, with one set of manual alignments chosen as nominally "correct," and the other set of alignments measured in relation to the first set.

Cosi et al. [32] reported on the manual alignment of 10 continuous-speech Italian sentences recorded at 16 kHz and aligned by three people. They found a mean deviation of 6 msec, about 55% agreement within 5 msec, and 93.5% agreement within 20 msec.

Ljolje et al. [92] reported on the manual alignment agreement for 100 Italian utterances from two human transcribers, and found 80.0% agreement within 10 msec, 92.9% agreement within 20 msec, and 96.8% agreement within 30 msec. These results correspond well with those reported by Cosi.

Wesenick and Kipp [147] evaluated the manual alignment of German sentences by three transcribers. They found average agreement levels of 63% within 0 msec (perfect correspondence), 73% within 5 msec, 87% within 10 msec, and 96% within 20 msec. The transcribers used in this study were all graduate students in phonetics, and all had received an intensive training session. As part of this training, a number of conventions were established to ensure consistent labeling. One such rule was to always set a segmentation boundary where the values of the speech signal changed from negative to positive [146]. Not surprisingly, these results represent the best reported performance of human consistency on the task of phonetic alignment.

Leung and Zue [89] evaluated 5 American English sentences as aligned by two people. The sentences were recorded using a microphone, and the text came from the Harvard list of phonetically-balanced sentences. Manual alignment required about 30 seconds per phoneme, and they reported approximately 80% agreement within 10 msec, 87% agreement within 15 msec, and 93% agreement within 20 msec.

Cole et al. [25] reported on inter-labeler agreement for four languages, as labeled by

both native and non-native speakers. For American English aligned by two transcribers (native speakers), they reported 79% agreement within 10 msec, which is marginally lower than the value reported by Leung. For German speech, they found 63% agreement within 5 msec and 79% within 10 msec when comparing two native-speaker labelers, and 69% agreement within 5 msec and 81% agreement within 10 msec when comparing a native-speaker labeler and a non-native-speaker labeler. One point of interest is that although Wesenick, Cosi, Ljolje, and Leung performed their evaluations on 16-kHz microphone speech and Cole et al. performed their evaluation on 8-kHz telephone-band speech, the results are quite comparable. In addition, the results for Leung on English speech and the results for Ljolje on Italian speech are nearly identical, as are the results for English and German alignments reported by Cole.

As none of the above evaluations were performed on the commonly-used TIMIT corpus of American English speech, we manually aligned 50 sentences from the test partition of TIMIT (1800 phonemes). We used the phoneme sequence as given in the TIMIT phoneme-label files, but removed all timing information prior to hand labeling. For evaluation, we (a) merged glottalized sounds with their surrounding voiced sounds, if possible, and (b) did not evaluate boundaries between stop closures and silence (as any such boundary is placed arbitrarily). We found 81.7% agreement with the standard TIMIT alignments with a threshold of 10 msec, and 93.5% agreement within 20 msec. These results correspond well with the results reported by Cosi, Ljolje, Leung, and Cole.

In summary, there is fairly consistent agreement among humans labelers for continuous speech, even across language and channel conditions. There is an average agreement of 93.78% within 20 msec for the measured manual alignments, with a maximum of 96% within 20 msec for highly-trained specialists using a set of rigorous and well-defined conventions.

3.2 HMM systems

As mentioned in Chapter 1, HMM and HMM/ANN speech recognizers can be used to obtain phonetic alignments using a process called forced alignment. In forced alignment,

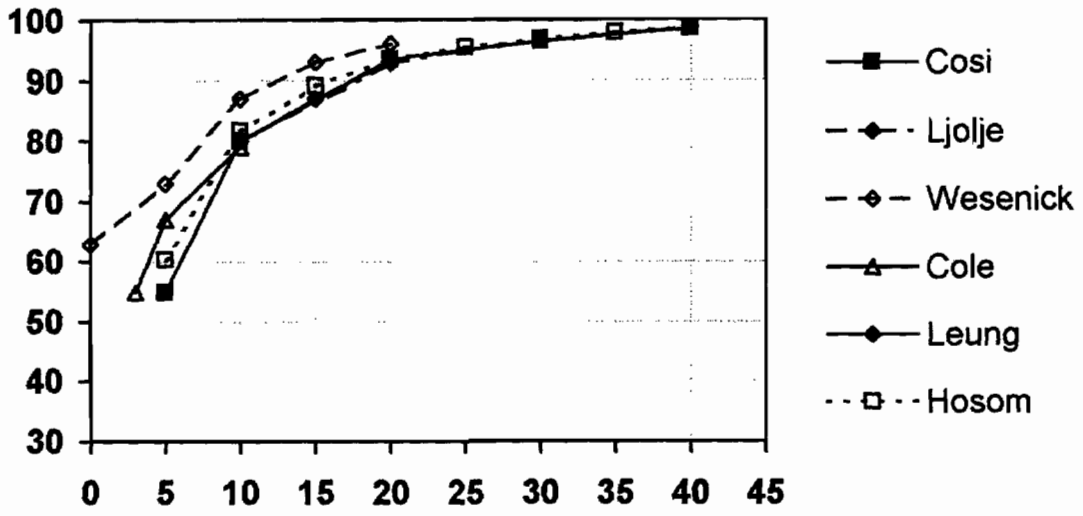

Figure 3.1: Inter-labeler agreement of alignments at various thresholds, as reported by six researchers.

the HMM is used to recognize the input speech with the Viterbi search constrained to only the correct sequence of phonemes. The result of the Viterbi search contains the phonetic alignment (as well as the score for the known phoneme sequence). In cases where the words are known but the phoneme sequence is not, a dictionary can be used in combination with pronunciation rules to generate a phoneme sequence for each word; these sequences can then be concatenated together, with optional pauses between words, to arrive at a phoneme sequence for the entire utterance. Rapp noted that because "the task of phoneme alignment can be considered as simplified speech recognition, it is natural to adopt a successful paradigm of [automatic speech recognition], namely HMMs, for alignment" [123].

Wightman and Talkin [149] developed an HMM-based system called "the Aligner," with the acoustic model training and Viterbi search implemented using the HTK Toolkit [150]. The Aligner uses a 10-msec frame rate and five mixture components per Gaussian to estimate the state occupation likelihoods. Non-speech sounds, such as breath noise

and lip smacks, are collapsed into a single "silence" model. The system was trained on unvoiced and voiced stop closures, whereas most HMM systems train the stop closure and the stop burst as one unit. The system was trained using the TIMIT labels as an initial segmentation. In evaluation of their system, they did not use the TIMIT phonetic sequence directly, but they first mapped the words to canonical dictionary pronunciations, then performed forced alignment, and then mapped the forced-alignment phonemes to the TIMIT phoneme sequence; this allowed them to compare the phonetic boundary alignments while still performing forced alignment from word-level information. Performance on the TIMIT test set was approximately 80% agreement within 20 msec.

Brugnara et al. [15, 16, 3] developed an HMM forced-alignment system that uses spectral variation features in addition to the standard cepstral-domain features for computing state occupation likelihoods. The use of these additional features resulted in a 2% relative reduction in error. They also tried adjusting the phonetic alignments after the Viterbi search, based on the values of the spectral variation features, but found no improvement in performance. They evaluated this system on the TIMIT database, and reported 75.3% agreement within 10 msec, 84.4% agreement within 15 msec, and 88.9% agreement within 20 msec. They also compared this system to an identical system trained without initialization from the manual alignment information, and found that the system trained using the manual alignments had a 50% reduction in error compared to the system trained without manual alignments.

Pellom [115, 114] developed an HMM for forced alignment with a variety of speech enhancement algorithms. This system uses a 5-msec frame rate, 5-state monophone HMMs, gender dependent models, 16 Gaussian mixture components per state, and Gamma distribution transition probabilities. When phoneme-level transcriptions are not available, the system generates pronunciations using the CMU dictionary and word-juncture modeling. The system was trained and evaluated on TIMIT data that had been down-sampled to 8 kHz, and agreement was 86.2% within 20 msec. Pellom evaluated the same system on the NTIMIT corpus of telephone-band speech and the CTIMIT corpus of cellular-band speech, using various noise-reduction techniques. For NTIMIT, the system with the best combination of speech enhancement algorithms had 76.8% agreement within 20 msec; for

CTIMIT, the best-performing system had 66.7% agreement within 20 msec.

Ljolje and Riley [93] built a three-state HMM system that has different types of phonetic models, depending on the availability of training data. If enough data are available for a given phoneme in its left and right contexts, then a complete triphone model is used, although the left and right contexts are clusters of similar phonemes instead of individual phonemes. If sufficient data are not available for a full triphone model, then a "quasi-triphone" model is attempted; this quasi-triphone model has the left state dependent on the left context, the middle state context independent, and the right state dependent on the right context. If sufficient data are not available for the "quasi-triphone" model, then left-context dependent and right-context dependent models are attempted. If sufficient data are still not available, then context-independent phoneme models are used. The HMM uses full-covariance Gaussian probability density functions to estimate the state occupation probabilities, a Gamma-distribution duration model, and a 10-msec frame rate. The models were trained and evaluated on the TIMIT database. Two types of models were trained: those based on the manual alignments in the TIMIT database, and those based on a mixture of manual alignments and Viterbi re-estimation of the alignments. In either case, they found 80% agreement within 15 msec.

Svendsen and Kvale [136] first segment the speech into acoustically similar segments, and then use an HMM with the segment boundaries as anchor points during the Viterbi search. Segmentation is done using vector quantization, constrained so that all vectors in a cluster are contiguous in time (called sequence-constrained VQ or SCVQ). Using this method, they set a threshold to provide 2.5 times as many segment boundaries as phonetic boundaries, so that 98% of the manually-labeled phonetic boundaries are within 20 msec of a hypothesized segment boundary. Then, a three-state monophone single-mixture HMM (with the ability to skip the middle state) is trained on each phoneme. During the Viterbi search, a state transition is only allowed at the hypothesized segment boundaries. This system was trained and evaluated on the EUROM0 corpus, which contains 16 kHz read speech from a small number of male and female speakers in different languages. Results on the single-speaker British-English test set showed 82.3% agreement within 20 msec, but this performance was probably negatively influenced by the lack of training data.

Cox, Brady, and Jackson [33] developed an HMM system with a 10-msec frame rate to align read British-English speech for 2 adult males, 2 adult females, and one female child. The system was trained on manually-segmented data and then evaluated on the same training data. This method of training and testing on the same data was acceptable for their purpose (generating alignments for only those five speakers), but does not allow comparison of their results with speaker-independent systems.

Ljolje, Hirschberg, and van Santen [92] trained a monophone (context-independent) three-state HMM system with Gaussian estimation of the state occupation likelihoods. Gamma distributions were used to model the phoneme durations, and the frame size was 2.5 msec. The system was trained using an initial uniform-duration segmentation of the states instead of manual alignments. Training and evaluation was done on Italian utterances in carrier phrases. When mean biases were removed from the results, performance was 78.1% agreement within 20 msec.

Pauws, Kamp, and Willems [113] trained an HMM system using a three-step process, so that they did not have to initialize their training with manual alignments. Their purpose was to create alignments for use in text-to-speech, and they wanted high-accuracy alignments without the costs associated with manual alignment. Their system was trained and evaluated on isolated Dutch words recorded from a single speaker. In the first step, the speech was segmented into three broad phonetic classes, "silence," "voiced," and "unvoiced," using energy in different frequency bands, the zero crossing rate, and the spectral slope. This initial segmentation had 82% agreement within 20 msec. Given this segmentation, the next step was to use sequence-constrained vector quantization (SCVQ) within each broad phonetic class to align the phonemes. This process resulted in 70% agreement within 20 msec. In the third step, an HMM system was trained to recognize each phoneme, with the initial segmentation taken from the second-step results. This system used 6 states per phoneme for all phonemes except bursts, which had 2 states per phoneme. The frame rate of 5 msec enforced a minimum duration of 30 msec for non-burst phonemes and 10 msec for bursts. Performance of this system was 89.5% agreement within 20 msec. Pauws, Kamp, and Willems then compared this system to a forced-alignment HMM system that

was initialized with manual segmentations, as well as another system that was initialized with equal-duration segmentations. Performance of the manual-segmentation system was 96.0% agreement within 20 msec, and performance of the equal-duration system was 76.14%. It should be noted, however, that the system trained with manual segmentations was trained and evaluated on the same data, so that this result can not be used for comparison with speaker-independent alignment systems.

Dalsgaard, Andersen, Barry, and Jörgensen [36, 37, 35] used a self-organizing neural network (SONN) to estimate the probabilities of distinctive phonetic features (such as "front," "central," "low," "high," and "labial," each with three possible values) in the speech signal. These distinctive features were subject to principle component analysis to determine the most relevant features for phonetic classification. The principle components were used to model phonetic likelihoods with Gaussian probability density functions, and then a Viterbi search was applied to these likelihoods to align the speech. Distinctive features were used so that the system could be easily applied to new languages. When evaluated on English speech from the EUROM0 corpus using 15 principle components, agreement was 77.1% within 20 msec.

Malfrere, Deroo, and Dutoit [96] compared alignments generated by dynamic time warping (DTW) of synthetic speech (TTS/DTW) with alignments generated by an HMM system. The HMM system was trained and evaluated on read French speech from a single speaker, and the system was initialized with the alignments from the TTS/DTW system (described below). The system was trained for several iterations using 16 Gaussian probability density functions per state. The alignments of this Gaussian system were then used to train a hybrid HMM/ANN system with a context window of nine frames. This final system had 84.0% agreement within 20 msec.

Kipp, Wesenick, and Schiel [75, 147] implemented an HMM system for use in cases where only the word-level transcription is available. This system performed simultaneous alignment of the canonical dictionary pronunciation and several pronunciation variants. The HMM system used context-independent models with between three and six states per phoneme, and a 10-msec frame rate. The HMM system was trained and evaluated on the PHONDAT-II corpus of German speech, and was initialized with manually-aligned data.

The post-processing refinement adjusted the boundaries within a 10-msec window using simple time-domain techniques. This system had 84% agreement within 20 msec.

In the approach developed by Stöber and Hess, a baseline approach similar to semi-continuous HMMs (where all states in the HMM share the same set of Gaussians, with different mixture component weights for each state) was augmented with specific duration information that had been scaled to fit the length of the utterance. The duration information predicted from the length of the utterance was modified based on a "fitness function," which was computed using a genetic algorithm. This approach yielded agreement of 84% within 20 msec on the Bonner Prosodische Datenbank (BPD) corpus, and agreement of 80% within 20 msec on the Phondat-II corpus.

Rapp [123] trained a forced-alignment system for German using Entropic's HMM Toolkit called HTK [150]. He used a 10-msec frame rate to report his results, but found that a 5-msec frame rate was also "acceptable." He reported 84.4% agreement within 20 msec for German read speech.

Wheatley, Doddington, et al. [148] trained an HMM on specific non-speech sounds (silence, inhalation, exhalation, and lip smack) as well as using gender-specific phoneme models. They trained and evaluated their system on telephone-band continuous speech. Evaluation was done by automatically determining the phoneme sequence of each word, and then comparing the word-level alignments of the automatic system with manual word-level alignments. They reported that their system had a "failure rate" of 0.9% (2 out of 212 sentences), and that "the overwhelming majority of words are correct at least to within a second; ... the alignment is normally correct to within one or two 20-msec frames." A lack of specific quantified results, and the fact that their system was trained and evaluated on telephone-band speech, makes a performance comparison with other systems difficult.

In summary, the reported systems represent numerous refinements on the standard HMM procedure, but in all cases the basic process remains the same, namely estimating phonetic likelihoods at each frame, and then searching through these likelihoods with a constrained Viterbi search to determine the phonetic alignments. Direct comparison of the results from these systems is not possible, because even in the four cases where the systems have been evaluated on the same corpus (TIMIT), there are small differences in

the implementation of the HMM systems that prevent a one-to-one comparison. In the case of Pellom's system, the TIMIT corpus was down-sampled to 8 kHz for training and evaluation, the frame rate was 5 msec, stop closures were merged with their succeeding plosives, and there were a total of 46 phonemes; in Brugnara's system, training was done at 16 kHz, the frame rate was 5 msec, stop closures were not merged with their succeeding plosives, and there were a total of 48 phonemes. Ljolje and Riley trained at 16 kHz with a 10-msec frame rate, merged stop closures with their bursts, and used a set of 47 phonemes. Wightman and Talkin trained at 16 kHz, used a 10-msec frame rate, did not merge stop closures, and used a set of 35 phonemes. If, however, we assume that the performance differences due to these variations are minimal (Wightman and Talkin claim "very similar" results for systems trained on 16 kHz and 8 kHz speech), then we can generally conclude that performance of HMM systems on the TIMIT database ranges from 80% to 88.9% agreement within 20 msec. Performance on other databases and languages tends to be similar but slightly lower, with agreement levels from 77% to 84% within 20 msec. Only Wheatley et al. and Pellom evaluated systems on telephone-band speech, and severe performance degradation was reported; even systems with the best possible noise compensation had no more than 76.8% agreement within 20 msec for land-line telephone speech and 66.7% agreement within 20 msec for cellular speech.

3.3 The DTW Approach to Phonetic Alignment

Dynamic Time Warping (DTW) is a method that aligns two sets of features in time, so that the error between the two features is minimized. It is a dynamic-programming algorithm, as is the Viterbi search; the Viterbi search, however, uses likelihood estimates and transition probabilities instead of a set of reference features and a distance metric. One of the earliest publications on using DTW for phonetic alignment of speech was published by Michael Wagner in 1981 [143]. This system is composed of the following stages:

1. LPC analysis is used to determine the energy, approximate formant values, degree of voicing, and fundamental frequency at each 5-msec frame. The voicing and fundamental frequency values are computed from autocorrelation of the inverse-filtered

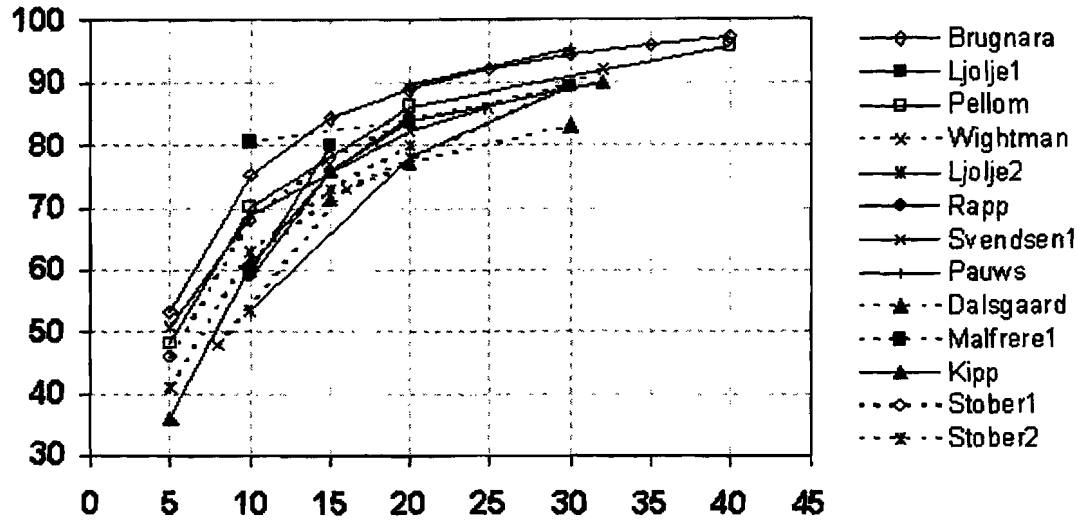

Figure 3.2: Agreement of HMM-based automatic alignment systems with manual alignments at various thresholds.

signal. (Methods of extracting voicing and fundamental frequency will be covered in more detail in Chapter 6).

2. The signal is classified into voiced, unvoiced, and silence segments by comparing the values from Step 1 to pre-determined thresholds. Segments of less than 20 msec are merged with surrounding segments.

3. Formant tracking of the voiced regions is done using points of expected reliability ("anchor points") in the signal.

4. The speech is aligned at the segment level using DTW to match the segments found in Step 2 (the "given sequence of segments") with the segments expected based on the phonetic transcription (the "expected series of segments"). A "distance table" of costs associated with mapping a given sequence of segments from the set {voiced, unvoiced, silence} to an expected sequence of segments from the set {vowel, unvoiced fricative, voiced fricative, unvoiced burst, voiced burst, nasal or semivowel,

stop closure, pause} is used, allowing the three types of segments from the acoustic analysis to be mapped to a set of eight broad phonetic categories.

5. The speech is aligned at the phoneme level, using DTW to match the phonetic transitions in the transcription with the phonetic transitions in the signal. The values for phonetic transitions in the signal are computed from the change in energy for the unvoiced regions and the change in formant values for the voiced regions; the values for phonetic transitions in the transcription are determined by table-lookup of "expected" energy changes and formant changes for the given phoneme pair.

This unique system, one of the first to use DTW and phonetic transition information in phonetic alignment, was reported to work "reliably" at the segment-alignment stage and "well" at the phonetic-alignment stage, based on a small corpus of two speakers. The author notes the advantage of using formant derivatives instead of absolute formant values, because "energy and formant derivatives are far more speaker-independent than absolute energies or formants" [143].

Malfrere, Deroo, and Dutoit [96] performed automatic alignment by dynamic time warping (DTW) of synthetic speech; this type of system will be referred to as a TTS/DTW system. This system generates the speech with the MBROLA speech synthesizer [43] (using a constant F0 value), and computes 36 spectral-domain features and energy values from the synthetic speech at each 10-msec frame. The same set of spectral and energy features is computed for the input speech. Finally, dynamic time warping is used to time-align the two utterances so that the differences (Euclidean squared distances) between the two sets of spectral features are minimized. The advantage of this approach is that no training database is needed; the disadvantage is that the synthetic speech from a single speaker serves as the only template. The TTS/DTW system was compared with an HMM alignment system that had been initialized using the output of this TTS/DTW system (described in more detail in Section 3.2). They evaluated the performance of both systems on data from a single French speaker. Results for the TTS/DTW system were 82.1% agreement within 20 msec; the HMM system attained agreement of 84.0% within 20 msec. Given the simple and constrained nature of the TTS/DTW system, and the

fact that the HMM system was trained using the output of the TTS/DTW system, it is interesting that the HMM system had only a 10% relative reduction in error.

Campbell [19] used a TTS/DTW system to align a corpus of Japanese speech. He used several TTS voices and prosodic contours to minimize the differences between the synthetic and input speech. This system had 69% agreement within 100 msec, which is far less accurate than his reported 97% agreement within 100 msec for two manually-generated alignments. Gong and Haton [58] performed phonetic alignment using the TTS/DTW approach, but performed an iterative process of alignment and speaker adaptation to minimize the differences between the synthetic and input speech. They evaluated their system by training two continuous-speech recognizers using the outputs of the baseline (first iteration) and final-iteration alignment systems, respectively. The recognizer trained on the final iteration's output had a 50% relative reduction in error, compared to the recognizer trained on the baseline system's output. Saito [129] proposed the use of the change in the fundamental frequency (delta-F0) for the alignment of speech. In Saito's system, DTW was used to align a new utterance with a reference utterance; then the boundaries obtained from DTW were adjusted based on the local maxima and minima in the delta-F0 contour. Saito reported average boundary-error reductions of 29% and 16% for two test speakers by incorporating the delta-F0 information into the alignment process. Svendsen and Soong [137] used DTW to align the input speech with "speaker-independent" phonetic templates obtained from spectral averages of different speakers. They reported agreement of 32% within 15 msec, 72% within 30 msec, and 92% within 45 msec. Falavigna and Omologo [45] also aligned the input speech with phonetic prototypes, but they used a spectral variation function to emphasize changes in the signal, and they also used the signal's energy contour to refine the DTW estimates. This system had 61% agreement within 20 msec when evaluated on an Italian continuous-speech corpus. Finally, Chamberlain and Bridle [20] modified the DTW algorithm for aligning two utterances, so that long samples of speech could be processed on low-memory machines.

In summary, DTW has been used, typically in conjunction with synthetic speech, to align the input utterance with a reference utterance. Although the methods of reporting performance are sometimes different than the standard method of percent agreement

within a given threshold, results in general do not seem as good as with HMM-based systems.

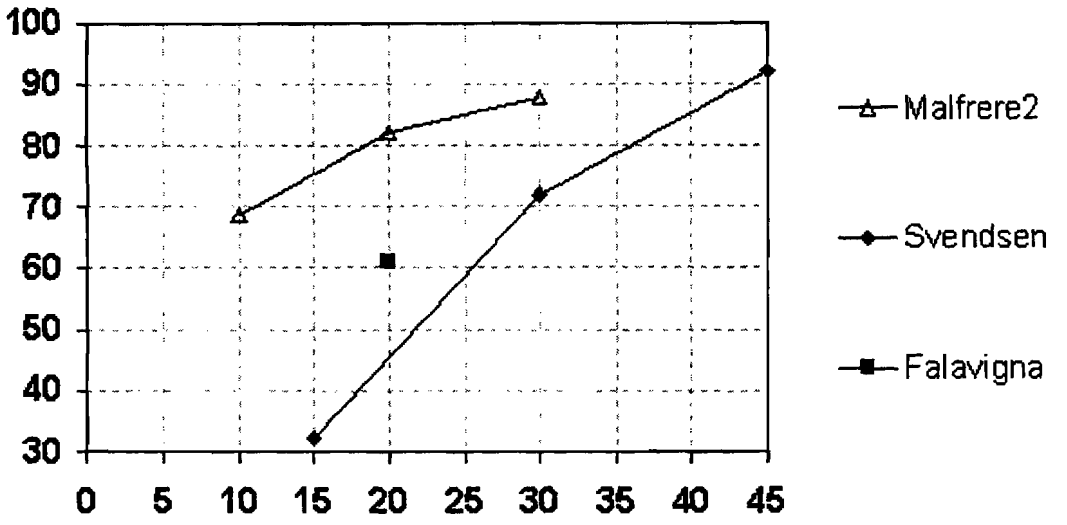

Figure 3.3: Agreement of DTW-based methods of automatic alignment with manual alignments at various thresholds.

3.4 Other Methods of Automatic Phonetic Alignment

The HMM and DTW approaches to phonetic alignment are certainly not the only ones; about one-third of phonetic-alignment systems reported in the literature use some other method. In this section, we describe some of these alternative systems, focusing on the ones that are more relevant to this thesis.

An alignment system developed by van Santen and Sproat [141] applies edge detectors to spectral-domain representations and energy information in different frequency bands. This information is combined with a set of phonetic sequences for each word to arrive at the aligned phonetic sequence. Their approach focuses on detecting phonetic boundaries (referred to as diphones) rather than the conventional HMM approach of estimating the likelihood of each phonetic category at every frame of speech. They note that the spectral

cues to different types of phonetic transitions are contained in different frequency bands; for example, a boundary between an /f/ and an /s/ has a decrease in energy below 2000 Hz and an increase in energy above 4000 Hz; a boundary between an /f/ and a vowel, however, has an increase in energy in the 800 Hz to 2500 Hz frequency range. The authors group the set of possible diphones into two classes, "broad" and "narrow." Broad diphones can be categorized by their manner of articulation (such as voiced burst or unvoiced fricative) and can be identified based on energy changes in broad regions of the spectrum. Narrow diphones are characterized by more subtle differences, such as formant movement or insertion of a glottal stop. To account for these two types of diphones, van Santen and Sproat use two representations of the speech signal; the first representation is energy in five different frequency bands (for classifying the broad diphones), and the second is a mel-frequency scale FFT representation (for classifying the narrow diphones). They then perform edge detection on the frequency bands, detecting both quick changes and less localized changes. The frequency-band information is combined in such a way that exact synchronicity in time is not required. This information is combined using Bayes' rule to estimate the "overall acoustic cost" of each diphone at each time frame. For the narrow diphones, the mel-FFT representation is used with vectors of weights that characterize each diphone to locate the time point of greatest change between the two phonemes. van Santen and Sproat reported 50% agreement within 2 msec and 95% agreement within 6 msec for a single speaker when evaluated on the training data, and 90% agreement within 20 msec when evaluated on a single test speaker. Although the use of a single speaker in the test corpus does not guarantee that the results will generalize to multi-speaker corpora, the extremely high agreement argues for the merit of this method.

Karjalainen, Altosaar, and Huttunen [73] also used boundary detection to automatically align speech; in their case, they used a set of neural networks. In a method similar to the work of van Santen and Sproat, they classify "coarse categories" of boundaries, such as "stop to vowel" and "vowel to nasal"; unlike van Santen and Sproat, they consider all boundaries to be "coarse," instead of splitting the boundaries into two classes. Using such coarse boundary detectors, they use 64 categories to cover the Finnish language. A simple rule-based parser is used to match the boundary-classification outputs with the text

transcription. The authors use LPC coding with a perceptually-warped frequency scale as input to the networks. These features, with a 100-msec context window and 10-msec frame size, are passed to a set of 64 feed-forward binary-output neural networks to classify the speech frames into the 64 boundary categories. The outputs of the neural network are smoothed, and smaller output peaks are removed. The outputs of the neural network are then matched to the transcriptions using a simple three-step rule-based algorithm. When evaluated on Finnish isolated words from a single speaker, recorded at 22 kHz, the mean alignment error on the test set was 8.7%; this compares favorably with an HMM-based alignment system trained on the same data, which had an error rate of 17.6%

Leung and Zue [89] developed a three-step procedure to align a phonetic transcription with its corresponding speech. In the first step, a classification of speech into broad phonetic categories is done, "to determine robust acoustic-phonetic events that are relatively context-independent." The classification is done using a binary decision tree; at each node in the tree, a binary decision is made based on features of the input signal. A 5-msec frame rate is used. At each frame, an M-dimensional vector is generated, where each element in the vector represents some acoustic-phonetic knowledge or information. The vector values are smoothed, clipped, and normalized. The speech is classified into a total of 6 categories: sonorant, obstruent, voiced obstruent, silence, nasal and voice-bar, and "unlabeled segments" that correspond to an energy dip in the middle of sonorant regions. Given this coarse segmentation, dynamic programming is used with acoustic-phonetic rules to map each coarse segment onto one or more phonemes. For example, a phoneme is not allowed to match a category outside of its class (obstruent can not match silence); also, a plosive such as /t^h/ is not allowed to match a long obstruent because of duration constraints. Then, to match segments that correspond to more than one phoneme, heuristic rules are used; for example, "pre- and post-vocalic liquids next to certain vowels are assumed to have a duration that constitutes one-third of the syllable nucleus." If clear acoustic cues are available, "further segmentation is accomplished by a proper selection of feature parameters and algorithms based on contextual information." This system was evaluated on three speakers reading a set of phonetically-balanced sentences which had been manually aligned. Results were approximately 75% agreement within 10 msec and 90% agreement

within 20 msec.

3.5 State of the Art in Phonetic Alignment

For the systems reviewed above that were evaluated on microphone-quality speech, performance ranged from 77% agreement [36] to 90% agreement [16, 89] within 20 msec; variables that may affect performance include the method of training the system, the number of speakers in the training and test corpora, the type of corpus (isolated word or continuous speech), and the language used in training and testing. Average performance is about 84% agreement within 20 msec, and HMM-based systems tend to out-perform other systems. If we focus on the TIMIT corpus of continuous English speech, then the reported performances range from about 80% [149] to 90% [16] within 20 msec, with an average agreement of 85% within 20 msec.

Pellom and Hansen [115, 114] were the only authors to train on one set of channel conditions and evaluate on a different set; this study showed performance on clean microphone speech of 86.2% within 20 msec, 76.8% within 20 msec for artificial telephone-band speech, and 66.7% within 20 msec for artificial cellular speech.

Manual alignment, in contrast, is reported to have inter-labeler agreement between 92.9% [92] and 96% [147] within 20 msec, with an average agreement of 93.78% within 20 msec. Manual-alignment agreement on the TIMIT corpus is 93.49% within 20 msec, which is a 41% reduction in error when compared to the best reported automatic results of 88.9% within 20 msec for TIMIT. The manual alignments of TIMIT are consistently more accurate than the automatic alignments across all thresholds, with a minimum error reduction of 25% within 5 msec, and with error reduction increasing steadily to 50% within 40 msec.

Figure 3.4 shows manual alignment agreement (the dotted line) on the TIMIT corpus as reported here, and automatic alignment performance on TIMIT (the solid line) as reported by Brugnara.

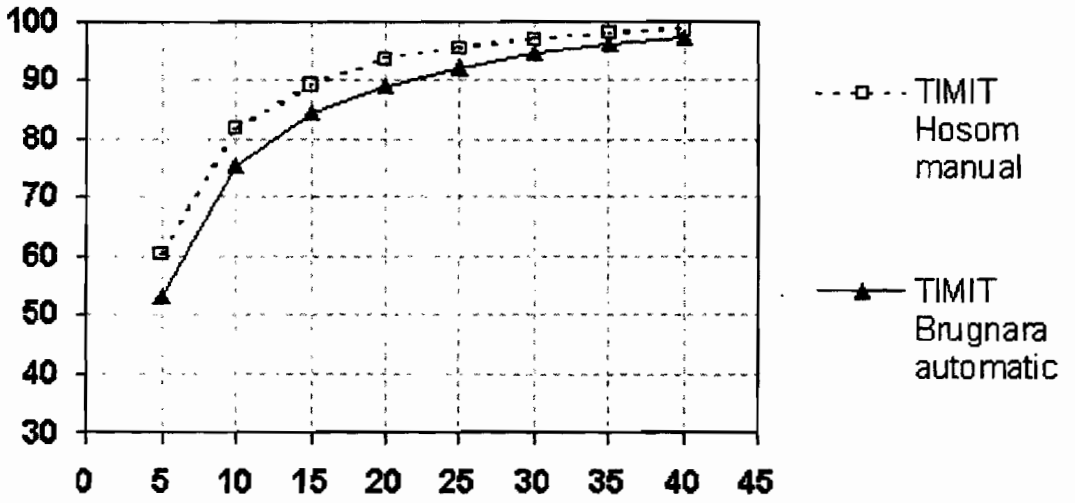

Figure 3.4: Comparison of manual alignment performance on TIMIT (dotted line), and best automatic alignment performance reported for TIMIT (solid line).

Chapter 4

Baseline System for Forced Alignment

In order to evaluate the method of phonetic alignment proposed in this thesis, it is necessary to have a baseline system that has been trained on the same data and evaluated using the same metrics. This section describes our baseline system, which was developed under the HMM/ANN framework commonly used at CSLU for speech recognition systems.

4.1 System Parameters

Most of the parameters used in training the baseline alignment system were determined from our experiments on training high-performance digit recognition systems [68]. These previous experiments investigated the use of PLP and MFCC features, delta features, the number of cepstral coefficients, the duration constraints used in the Viterbi search, and the type of context-dependent categories used in recognition. The set of "best" parameters was obtained based on word-level results from a large set of parameter combinations.

Training was done on the TIMIT corpus [52] (for samples of microphone speech), the OGI Stories corpus [27] (for samples of telephone-band speech), and the OGI Portland Cellular corpus [27] (for samples of telephone-band cellular speech). This variety of corpora was used to make the system robust to several channel conditions, instead of being specific to one type of channel.

The TIMIT corpus (a joint effort between MIT, Texas Instruments, and SRI) contains read speech from 630 speakers from eight dialect regions of the United States. The sentences were designed to be phonetically rich, and were recorded with a Sennheiser noise-canceling, head-mounted microphone in a quiet environment. The speech was digitized at

16 kHz with 16-bit resolution. The corpus contains waveform data, text transcriptions, and time-aligned phonetic labels.

The OGI Stories corpus contains utterances of extemporaneous speech, where each utterance is approximately 50 seconds in length. These data were recorded over telephone channels. Speakers were recruited from throughout the United States, and were asked to speak on the topic of their choice for one minute. A total of 692 utterances were recorded, of which more than 200 have been transcribed with time-aligned phonetic labels. The data were recorded from an analog line, and digitized in 8 kHz 16-bit linear format.

The Portland Cellular corpus consists of utterances obtained from speakers who were using cellular telephones. Like the Stories corpus, the Portland Cellular corpus contains extemporaneous speech on a topic of the speaker's choice, and 200 calls have been transcribed at the phonetic level. The data were captured digitally from a T1 connection and saved in 8 kHz, 8-bit μ-law format.

One of the first steps in training the baseline system was to automatically map the hand-labeled phonetic symbols in the training corpora to a consistent set of symbols suitable for training. This mapping consisted of removing diacritic symbols, mapping the phonetic symbols to the Worldbet system, if necessary, and mapping non-speech labels to the silence (/.pau/) label. In addition, very short pauses (with duration less than 20 msec) were removed to improve the number of available contexts, and glottalization labels were merged into the neighboring vowel (or sonorant) or split between surrounding vowels.

The sub-phonetic categories in the baseline system are similar to those used by Ljolje and Riley. Each phoneme is split into one, two, or three sub-phonetic parts. If split into two or three parts, the left part is dependent on the context of the preceding phoneme's broad category, the center part (if any) is context independent, and the right part is dependent on the following phoneme's broad category. Phonemes that remain as a one-part phoneme can either be context-independent (for example, the characteristics of /.pau/ do not depend on either the preceding or following phoneme) or dependent on the following phoneme (for example, burst sounds such as /t^h/ have characteristics that depend on the following vowel, but are always preceded by silence). The left and right contexts for each category are not phoneme-specific, but contain clusters of phonemes grouped into a

common broad category. The phoneme clusters that comprise each context are specific to each target phoneme, and each cluster of phonemes was determined using a tree-based clustering procedure.

Transition probabilities for each state were set to be all equally likely, so that the implicit Geometric distribution found in standard HMM systems was removed. To make use of *a priori* information about phonetic durations, the search was constrained by specifying minimum and maximum duration values of each category. The minimum duration for a category was set to be the value at the 2nd percentile of all duration values for that category, and the maximum duration was set to be the longest duration for that category found in the training data. Percentile values were used instead of the absolute minimum durations to remove outliers. During the search, hypothesized category durations beyond the minimum or maximum value were penalized by a value proportional to the difference between the proposed duration and the specified minimum or maximum duration.

The speech data were converted to a 16 kHz sampling rate, if necessary. A 160 Hz high-pass filter was applied to make the microphone-speech training data more closely match the telephone-speech data, and to remove low-frequency breath noise that is sometimes present in microphone speech. The system was trained using 13 PLP features (12 cepstral coefficients and 1 energy parameter) as well as their delta values. A window size of 16 msec was used with a frame rate of 5 msec. Cepstral-mean subtraction (CMS) was used to reduce convolutional noise.

As many as 8000 samples of each sub-phonetic category were taken from each corpus, for a total of up to 24,000 training samples per category. A context window of 5 frames was used, with frames taken at –60, –30, 0, 30, and 60 msec relative to the frame of interest. The resulting set of 130 features was input to a fully-connected feed-forward neural network, which was trained using back-propagation to estimate the likelihood of each sub-phonetic category. The training was adjusted to use the negative penalty modification proposed by Wei and van Vuuren [145] instead of division by priors. The network had 130 inputs, 300 units in the hidden layer, and 614 outputs. A total of nearly 2 million examples were used during training, corresponding to about 1 GB of data. Training was done for 45 iterations. The network results for iterations 15 through 45 were then applied

Table 4.1: Performance of the baseline alignment system on the TIMIT, NTIMIT, and CTIMIT corpora. The thresholds for agreement are specified in each column heading.

Corpus	5 msec	10 msec	15 msec	20 msec	25 msec	30 msec	35 msec
TIMIT	48.23	72.95	84.13	89.95	93.21	95.21	96.47
NTIMIT	42.52	65.12	76.80	83.23	87.43	90.20	92.18
CTIMIT	30.30	48.64	60.90	68.20	73.38	77.07	79.87

to the forced-alignment task on a development set of the TIMIT corpus, and the "best" iteration was determined by selecting the iteration with the minimum alignment error.

4.2 Performance

This baseline system was evaluated on the TIMIT, NTIMIT, CTIMIT, Stories, and Portland Cellular corpora. The baseline method has 89.95% agreement within 20 msec on TIMIT, which is a 9% reduction in error compared to the best reported automatic-alignment results on this corpus [16]. However, the manual inter-labeler agreement of 93.49% within 20 msec on TIMIT is still a 35% reduction in error compared to this baseline. For the NTIMIT corpus, the baseline system has 83.23% agreement within 20 msec. The inter-labeler agreement of 89.66% on NTIMIT thus represents a 38% reduction in error over the baseline system. When evaluated on the CTIMIT corpus, the baseline system has agreement of 68.20% within 20 msec; the inter-labeler agreement of 80.74% is a 39% reduction in error over the baseline performance. The baseline system has 87.35% agreement within 20 msec on the Stories corpus, and 82.51% agreement within 20 msec on the Portland cellular corpus, indicating that the artifical means of creating the NTIMIT and CTIMIT corpora by sending the TIMIT data through telephone and cellular channels does not create data that are representative of actual telephone or cellular speech. The performance of the baseline system on the TIMIT, NTIMIT, and CTIMIT corpora at several thresholds is given in Table 4.1, and the level of inter-labeler agreement on these corpora is given in Table 4.2.

The performance of this baseline system, the system in the literature with the best

Table 4.2: Level of inter-labeler agreement on the TIMIT, NTIMIT, and CTIMIT corpora. The thresholds for agreement are specified in each column heading.

Corpus	5 msec	10 msec	15 msec	20 msec	25 msec	30 msec	35 msec
TIMIT	60.38	81.73	89.07	93.49	95.36	96.91	97.79
NTIMIT	46.23	71.94	84.29	89.66	92.34	93.91	95.14
CTIMIT	44.51	61.87	73.61	80.74	84.65	87.92	91.03

reported results on TIMIT, and manual alignments as evaluated on TIMIT are plotted in Figure 4.1. The performance of this system as compared to manual alignments on the NTIMIT and CTIMIT corpora are shown in Figures 4.2 and 4.3, respectively. (The levels of manual agreement on the TIMIT, NTIMIT, and CTIMIT corpora were obtained by manually aligning 50 randomly-selected sentences from each corpus, and comparing these alignment results with the canonical TIMIT alignments).

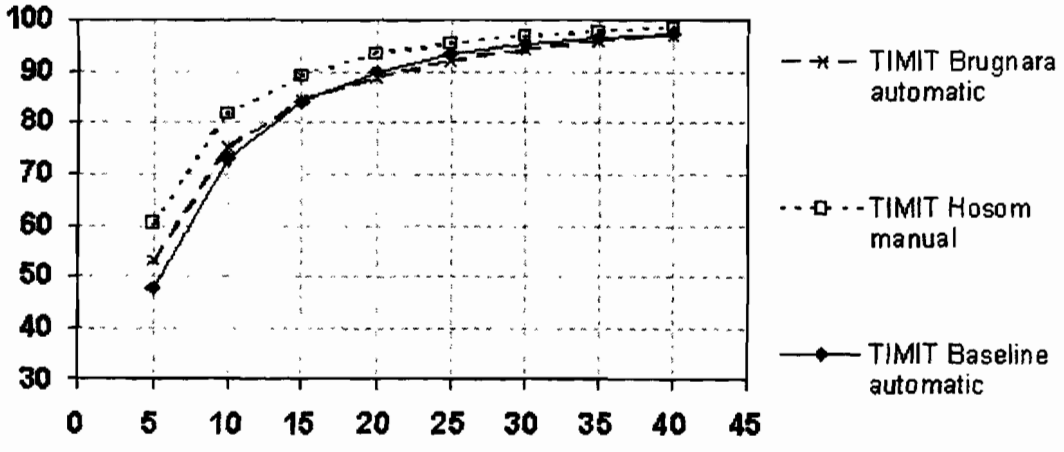

Figure 4.1: Performance on the TIMIT test set for manual alignments (dashed line), best reported results (by Brugnara et al., dotted line), and the baseline system used in this thesis (solid line).

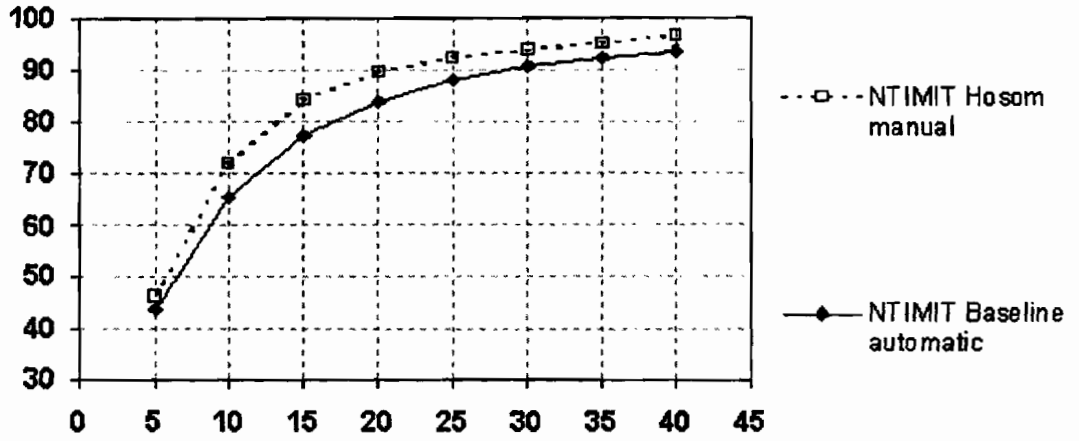

Figure 4.2: Performance of manual labeling (dotted line) and baseline system labeling (solid line) on the NTIMIT corpus.

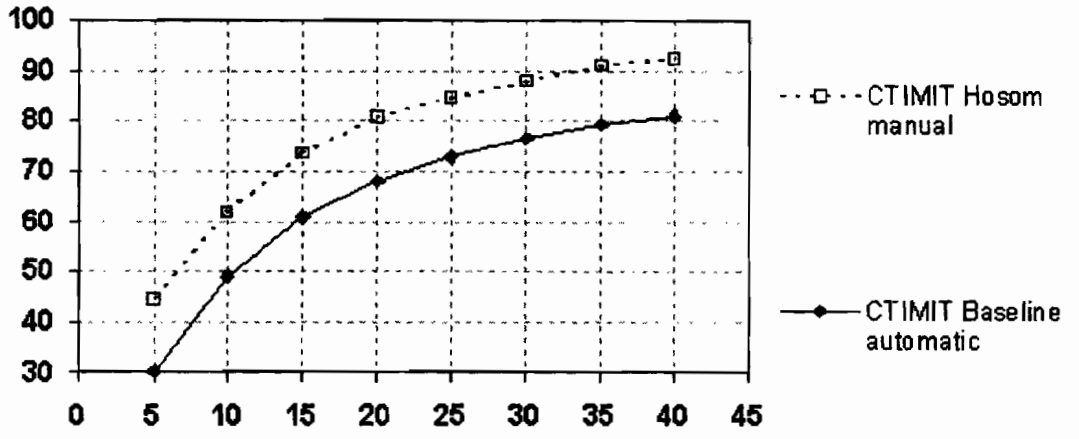

Figure 4.3: Performance of manual labeling (dotted line) and baseline system labeling (solid line) on the CTIMIT corpus.

Chapter 5

Proposed Approach

With performance of state-of-the-art forced alignment being between 35% and 40% worse than observed manual alignments on TIMIT, NTIMIT, and CTIMIT, it is clear that performance improvement should be possible. Our approach to realizing performance gains is to understand and improve upon the weaknesses of current HMM/ANN systems. We continue to rely on the HMM/ANN model as a foundation, because of its solid framework and superior results when compared to other systems. However, given our knowledge of human speech production and recognition, we believe that the HMM/ANN model can be improved to incorporate more of the information that is used by humans when recognizing speech, thereby bringing performance closer to that of human levels. Our hypothesis is that the integration of such acoustic-phonetic information into an HMM/ANN alignment system will significantly improve its agreement with manual alignments and its robustness. We base our thesis on the assumptions that the multiple-cue model of speech is valid, that invariant cues can be identified, and that special, inherently-complex decoders (as in Liberman's motor theory) are not required for automatic speech recognition.

The proposed model addresses the integration of acoustic-phonetic information from three angles: with the integration of acoustic-level features, with the use of phonetic transition information, and with the use of distinctive phonetic features. For the acoustic-level features, we supplement the current HMM/ANN spectral-domain features with specific acoustic-phonetic features believed to be important for speech perception. For phonetic transition information, we identify not only the context-dependent sub-phonetic categories of current HMM/ANN systems, but we simultaneously identify and integrate phonetic transition categories. For the distinctive phonetic features, we combine distinctive feature

information representing phonetic manner, place, and height, to arrive at a phoneme-level representation. Each of these levels will be described in more detail in this chapter and the following chapters.

In the implementation of this approach, we have assumed that the correct phonetic sequence of each word is known, in order to separate word-level influences from phoneme-level alignment performance.

5.1 Acoustic-Level Features

Current HMM/ANN systems use features that represent spectral information, energy, their delta values, and possibly their delta-delta values (acceleration coefficients) at each time frame. Usually the spectral information is warped to emphasize perceptually relevant aspects, and either cepstral-mean-subtraction (CMS) or RASTA processing is used to attenuate convolutional (channel) noise. This feature set represents a generic view of the speech signal with values that are useful for classification of all phonemes. However, such features do not give a complete representation of all relevant information in the speech signal. For example, knowledge that a frame of speech is voiced is clear evidence that the corresponding phoneme can not be a voiceless fricative or affricate; a frame with vowel-like spectral characteristics but a lack of voicing is more likely to be the consonant /h/ than if voicing were present. Also, knowledge that a frame of speech is glottalized increases the likelihood that the frame is at the beginning or ending of a word. The standard feature set, however, does not capture explicit information about voicing or glottalization. These types of features will be referred to here as "acoustic-level features," to distinguish them from other types of features such as phonetic transitions or distinctive phonetic features (described below in Sections 5.2 and 5.3). From the results of the research by Liberman as well as Cole and Scott, it is believed that humans use as many relevant cues as possible, which motivates us to incorporate acoustic-level features that are complementary to the standard feature set.

Perceptually-relevant features, such as voicing and formants, have been used before in speech recognition, but without dramatic success [65, 130]. Improvement has been

limited because of insufficient accuracy in their extraction. Experiments that use the "correct" values for voicing, broad-category features, or formants (as determined by manual assignment or from the phonetic transcription), instead of values extracted from the speech signal, have demonstrated at least twice the amount of error reduction [65, 130]. As a result of the discrepancy between the theoretically-possible results and the actually obtained results from extracted parameters, we focus on developing methods for robust extraction of perceptually-relevant features.

Several criteria were used in selecting the features for investigation. First, we focused on features that are not well represented by the standard spectral-domain features; we assumed that if the standard feature set implicitly captures the information of interest (such as formant movement), then the neural-network classifier will not benefit greatly from redundant information supplied in a different form. Second, as models of phoneme duration, coarticulation, and other time-domain aspects of speech are still in preliminary stages [79, 22], we restricted our investigation to features that are local in time and can be determined using fixed time windows. Third, we applied our knowledge of speech production, human speech recognition, and automatic speech-recognition systems to identify features that may provide relevant information about phonetic identity. Fourth, extraction of a feature from the speech signal must be computationally tractable, as the final alignment system is required to be significantly faster than manual alignment.

Based on these criteria, we developed a list of five features that merited further investigation. These features are intensity discrimination, voicing, fundamental frequency (F0), glottalization, and burst-related impulses.

Intensity Discrimination. We implement a measure of intensity discrimination in automatic speech recognition. Intensity discrimination has been modeled in psychological studies as a relative change in energy on the log scale [102]. This model has been found to provide a good description of human detection of intensity changes. As changes in intensity provide useful information about phonetic transitions, a perceptually-motivated model of intensity discrimination may be useful in automatic phonetic alignment.

Voicing. Voicing is a measure of periodicity in the waveform that occurs when the vocal folds vibrate. Voicing information can be used to distinguish between phonemes (such as /s/ and /z/, or /h/ and a vowel); the time from the current frame until the onset of voicing (voice-onset time) also provides information about the identity of plosive consonants (/p^h/, /t^h/, /k^h/, /b/, /d/, /g/, /ʧ/, and /ʤ/). Extracting information about voicing is a well-researched area, and yet a definitive, reliable method does not exist. Voicing extraction is particularly difficult on telephone-band speech, in which the lower regions of the frequency spectrum (which may contain the first few harmonics related to voicing) are severely attenuated.

Fundamental Frequency. Fundamental frequency, or F0, is the rate at which the vocal folds vibrate during voiced speech. As Saito noted [129], a change in F0 may indicate a phonetic boundary between voiced consonants and vowels. Methods of extracting F0 are also quite numerous, but again a definitive, robust method is still an area of research, and F0 extraction on telephone-band speech is considered a difficult topic.

Glottalization. Glottalization is defined as aperiodic or extremely slow vibration of the vocal folds, which sometimes occurs at word boundaries. Glottalization may be the only cue that identifies a word boundary, if the spectral characteristics of the ending phoneme of the first word and the beginning phoneme of the second word are the same. (Examples of this can be seen in the words "heavy yoke" and "E.E.")

Impulses. Impulses are defined here as the sudden increase in energy that occurs at the beginning of a burst. Identification of impulses helps to identify plosive consonants and locate their initial boundary. There are some previously-published methods for identifying bursts, and we will review the literature and propose and evaluate a perceptually-motivated method.

The feature set we consider for this thesis is not a complete set, in that it does not attempt to represent all of the features that are believed to be used by humans during speech recognition, nor all of the possible features that could be used. The set is, however, composed of features that are thought to be relevant to speech perception, are not well

represented by current spectral-domain features, have well-researched models, do not rely on higher-level phonetic knowledge, and are computationally tractable.

Integration of these features into a speech recognition system can be accomplished by early integration (during phonetic classification by the ANN or GMM), middle integration (during the Viterbi search), or late integration (by performing post-Viterbi combination of word-level outputs). Early integration has the advantage that decisions are made while all of the relevant information is available and can be combined in a non-linear way. As the neural network can easily take any fixed number of values as input and arrive at a theoretically optimal classification of the input space, we integrate the proposed additional features into the alignment system by appending them to the existing set of PLP features for input to the neural network.

The contribution of this thesis to the area of feature extraction for phonetic alignment is in the development of new methods for robust extraction of acoustic-level features.

5.2 Phonetic Transition Information

5.2.1 Motivations for Phonetic Transitions

In most HMM and hybrid HMM/ANN systems, the categories for recognition are context-dependent sub-phonetic units that are trained on all frames within a sub-phonetic segment. For example, in our baseline system, a category can correspond to a whole, half, or third of a phoneme, and the left and right sub-phonetic categories are dependent on the preceding and following phonetic contexts, respectively. Each state in the HMM is then associated with a single phonetic-based category, and the state transition information is determined using *a priori* information in the training set. As a result of this framework, the likelihood of a state transition is not dependent on information in the speech signal being recognized. Furthermore, classification results may be less reliable at phonetic boundaries, because the speech signal has a higher degree of variability during a transition, there are fewer transition examples than non-transition examples for training, and the data are more widely spread in the feature space due to coarticulation between the

two phonemes. However, perceptual studies by Furui [51], the analysis of spectrogram-reading techniques by Cole [29], the (sometimes preliminary) speech-recognition systems developed by Zue and Glass, Cravero, Morgan, and Hosom [56, 34, 106, 67], and the phonetic alignment system developed by van Santen and Sproat [141] indicate that phonetic boundaries are well-motivated categories for classification, providing useful information that is complementary to the phonetic steady-state regions. The development of an HMM system that incorporates acoustics-dependent phonetic transition information may then lead to more robust recognition and alignment.

5.2.2 Previous Approaches to Phonetic Transitions

In the MIT SUMMIT system [56], the likelihoods of "events" are computed, where an event corresponds to a significant change in the signal acoustics. Given a particular segmentation from a segment-based recognizer, each event may be considered either a phonetic boundary or internal to a phoneme. The likelihood of each observed event's acoustics is then determined, and the total likelihood of the series of events in the segmentation is computed by multiplication. This likelihood can be integrated with the segment-based phoneme likelihood by assuming independence and multiplying the values.

In a system developed at OGI [67], a hybrid HMM/ANN system classifies speech into context-independent phonemes or phonetic transition regions (modified diphones). The diphone units used in training are up to 120 msec in length; if a phoneme in the training set is longer than 120 msec, then a context-independent phonetic category is created in the middle of that phoneme. During the Viterbi search, the context-independent steady-state region is made optional, in order to account for rapid speech. In this approach, the probabilities of steady-state and transition categories are considered independent, just as all categories are considered independent in a standard HMM system. However, a relationship between these two types of categories is enforced during the Viterbi search, namely that steady-state and transition categories must occur in alternating sequence.

In a system developed at CSELT [34], the units for classification are context independent phonemes and phonetic transitions, as in the OGI system. A major difference is that the CSELT system does not model every possible phonetic transition, but only the ones

that are thought to be perceptually important. Similarly, some phonemes (such as plosives) are not modeled by context-independent units. This approach reduces the number of categories in a general-purpose recognizer to 123 (22 for context-independent phonemes, and 101 for phonetic transitions). A second difference is that in the OGI system, phonetic transition regions are allowed to occur over a variable number of frames, whereas in the CSELT system, the phonetic transitions are restricted to a contiguous series of four frames (40 msec).

In the Stochastic Perceptual Auditory-event-based Model (SPAM) approach [106], transition-based recognition and phonetic-category recognition are performed independently and then combined after each Viterbi search. The transition-based recognizer has phonetic-transition categories and a single "non-transitioning state" ("nts") category. The transition categories are associated with states that have no self-loop, and these states are separated by the "nts" state which does have a self-loop. This recognizer and a standard context-independent phonetic-category recognizer are run separately, and the Viterbi-search results from each recognizer are combined to obtain a single score for each word. Although this approach overcomes the weakness of using a single classifier for both phonetic-category and transition recognition, it is similar to the SUMMIT system in that the transition and phonetic-category information are combined at a late stage in the recognition process.

In the theoretical domain, Bourlard and Morgan proposed Discriminant HMMs, in which the likelihood of a state is estimated given an observation of speech data. This is in contrast to the standard HMM approach, in which the likelihood of an observation is estimated given a particular state. In Discriminant HMMs, the likelihood of observing a given state is dependent not only on the observation vector, but also on the prior state, which can be represented as

$$p(q_n^m | q_{n-1}^k, x_n, \theta)$$

where q_n^m is state m at time n; q_{n-1}^k is state k at the previous time $n-1$, x_n is the observation data, and θ is the set of model parameters. This dependence of the current state on the previous state is related to the topic of state transitions, in that in both cases the relationship between the previous state, the current state, and the observed

speech must be learned. Although this technique is appealing, they reported that "facing numerous problems, this approach was however simplified by ... disregarding the previous state in the conditional." Division by the state priors then reduces this technique to the more conventional HMM framework.

Bengio proposed the use of asynchronous Input-Output HMMs (called asynchronous IOHMMs or just IOHMMs) to the problem of speech recognition [8]. In standard HMMs, the speech observations are considered outputs (states "generate" observations), and the state sequence that best matches this known output is computed. In IOHMMs, in contrast, the speech observations are considered the input and the phoneme sequence is the output. Both the state emission probabilities and the state transition probabilities are dependent on the observed speech input. For the task of speech recognition, in which the input sequence of speech observations is generally longer than the output sequence of phonemes, there is an additional "emit-or-not distribution"; when a state is entered, a decision is made whether or not to emit a phoneme output based on this emit-or-not distribution. The stated potential advantages of IOHMMs over conventional HMMs include the following:

(a) Training is discriminant (which is also true for hybrid HMM/ANN systems),

(b) The emission and transition distributions can be modeled using ANNs (whereas in conventional HMM/ANN hybrids, only the emission distributions are modeled by ANNs), and

(c) The reduction from the large number of possible outputs in standard HMMs (the number of possible output observations) to the smaller number of possible outputs in IOHMMs (the number of phonemes) "reduces the problem of imbalance between transition probabilities and emission probabilities."

We are not aware of any implementations of IOHMMs for speech recognition, which limits discussion to the theoretical domain.

Finally, Riis and Krogh [127] proposed an approach called "hidden neural networks" (HNNs); as part of this approach, neural networks can be used to independently estimate the probability of state occupation as well as the probability of state transition. Riis and

Krogh give no details about how the transition networks should be implemented, but they report that they "did not observe any improvements using transition networks," and so in their final system all models "use standard HMM transition probabilities." As a result, their final system is similar in many ways to more standard HMM/ANN systems, although the method of training the ANNs differs.

5.2.3 Proposed Approach to Phonetic Transitions

Our proposed approach to integrating transition information is motivated by the importance of phonetic transitions and by the desire to integrate this information with phonetic-category information early in the decision-making process. In standard HMMs, the probability of an observation sequence given a state sequence is defined as the multiplication of the likelihoods of each observation at each state:

$$P(\mathbf{O}|\mathbf{q}) = \prod_{t=1}^{T} p(\mathbf{o}_t|q_t) \tag{5.1}$$

where $\mathbf{O}$ is the observation sequence $(\mathbf{o}_1\mathbf{o}_2\mathbf{o}_3 \ldots \mathbf{o}_T)$ from time 1 to time T and $\mathbf{q}$ is the state sequence $(q_1q_2q_3 \ldots q_T)$. In our system, we define the probability of an observation sequence given a state sequence to be the multiplication of the likelihoods of each observation given the current state and the transition from the previous state to the current state:

$$P(\mathbf{O}|\mathbf{q}) = p(\mathbf{o}_1|q_1) \cdot \prod_{t=2}^{T} p(\mathbf{o}_t|\text{trans}(q_{t-1}, q_t), q_t) \tag{5.2}$$

where $\text{trans}(q_{t-1}, q_t)$ represents the transition from state q_{t-1} to state q_t. In the case where q_{t-1} is the same state as q_t, $\text{trans}(q_{t-1}, q_t)$ represents the probability of a self-loop, independent of which state is currently occupied. We can then use Bayes' rule to transform this so that the state information is dependent on the observation vector:

$$p(\mathbf{o}_t|\text{trans}(q_{t-1}, q_t), q_t) = \frac{p(\text{trans}(q_{t-1}, q_t), q_t|\mathbf{o}_t) \cdot p(\mathbf{o}_t)}{p(\text{trans}(q_{t-1}, q_t), q_t)} \tag{5.3}$$

We then assume independence between the state transition and the state occupation probabilities:

$$p(\mathbf{o}_t|\text{trans}(q_{t-1}, q_t), q_t) = \frac{p(\text{trans}(q_{t-1}, q_t)|\mathbf{o}_t) \cdot p(q_t|\mathbf{o}_t) \cdot p(\mathbf{o}_t)}{p(\text{trans}(q_{t-1}, q_t)) \cdot p(q_t)} \tag{5.4}$$

and factor this into three separate parts:

$$p(\mathbf{o}_t|\text{trans}(q_{t-1},q_t),q_t) = \frac{p(\text{trans}(q_{t-1},q_t)|\mathbf{o}_t)}{p(\text{trans}(q_{t-1},q_t))} \cdot \frac{p(q_t|\mathbf{o}_t)}{p(q_t)} \cdot p(\mathbf{o}_t) \tag{5.5}$$

Finally, noting that a neural network estimates the probability of each state given an observation vector, and that our training has been modified to implicitly divide by the prior probability of the state, we obtain

$$p(\mathbf{o}_t|\text{trans}(q_{t-1},q_t),q_t) = \text{ANNtrans}_{q_{t-1},q_t}(\mathbf{o}_t) \cdot \text{ANNphon}_{q_t}(\mathbf{o}_t) \cdot p(\mathbf{o}_t) \tag{5.6}$$

where $\text{ANNtrans}_{q_{t-1},q_t}(\mathbf{o}_t)$ is the neural-network estimate of the state transition probability from state q_{t-1} to state q_t given the observation $\mathbf{o}_t$, and $\text{ANNphon}_{q_t}(\mathbf{o}_t)$ is the neural-network estimate of the phonetic-category probability at state q_t given the observation. (For $\text{ANNtrans}_{q_{t-1},q_t}(\mathbf{o}_t)$, if the phoneme corresponding to q_t is equal to the phoneme of state q_{t-1}, then the estimate is of the likelihood of remaining in the same phoneme, regardless of the identity of the particular state.) Because $p(\mathbf{o}_t)$ is constant for any given utterance, the multiplication of the neural-network outputs is a scaled estimate of the likelihood of the observation sequence given the state sequence.

This framework allows us to include, for each observation, information about the state occupation and state transition likelihoods. We construct two separate networks, one each for estimating transition probabilities and phonetic-category probabilities. The transition classifier estimates the probability of phonetic transitions, with a single "non-transitioning state" (as in the SPAM model) to estimate the probability of a self-loop or phoneme-internal transition. For the work done in this thesis, the transition classifier depends on the use of distinctive phonetic features, which will be discussed below, and so a discussion of the implementation details will be postponed until Chapter 7

In summary, the transition-based systems that have been implemented perform late integration of phonetic category and phonetic transition probabilities, or use a single classifier to estimate both (context-independent) phonetic steady-state and phonetic transition probabilities. The theoretical approaches that have been proposed have, upon implementation, not included transition probabilities that are dependent on the input observations.

The proposed approach combines the phonetic category and phonetic transition information at an early stage in the classification process, trains separate classifiers for phoneme-based and transition-based recognition (thereby making context-dependent modeling of the phonetic categories more practical), allows discriminative training of the transition probabilities, and is computationally tractable.

5.3 Distinctive Phonetic Features

5.3.1 Motivations for Distinctive Phonetic Features

In most HMM and HMM/ANN systems, the basic unit of recognition is the phoneme. According to linguistic theory, each phoneme can be further decomposed into some number of independent and distinctive features; the combination of these features serves to uniquely identify each phoneme. The use of distinctive features in phonetics, in the TRACE model, and in spectrogram reading suggests that these features may also be advantageous in approaches to automatic speech recognition. The motivations for using distinctive phonetic features are varied, and include the belief that distinctive features will "minimize extra-linguistic information" such as speaker variability and signal distortions [82, 77], or that they will provide better modeling of coarticulation [77]. Another motivation is based on belief in Liberman's motor theory, which postulates that phonetic recognition is only possible through the identification of physical motor gestures, and such gestures may be closely related to, or mapped to, distinctive features.

In our case, the primary motivation for using distinctive phonetic features is to increase the amount of training data per node in the neural-network. By training on sets of independent distinctive phonetic features, each with a smaller number of categories than the context-dependent phonetic and phonetic-transition categories, and by then combining the probabilities of these (independent) categories according to perceptually-motivated rules, the networks can be trained with more data per node. A second motivation is to allow our proposed system to be easily extended to other languages; for example, we hope that by the addition of a distinctive feature such as lip rounding, we can gracefully extend our system that has been trained on English speech to process German speech as well.

If a phonetic-based system were used, the network would have to be entirely re-trained. A third motivation is to enable research on pronunciation modeling. For example, the final vowel in the word "seven" can be pronounced in many ways; instead of mapping the wide variety of pronunciations to a single quasi-phonetic unit (as is done currently), and instead of providing a large number of alternative pronunciations in the dictionary, a more elegant solution may involve specification of a word in terms of features that uniquely specify that word. In the case of "seven" in a digit-recognition system, nearly any vowel is permissible, and with the use of distinctive phonetic features, it may be easily possible to specify an inclusive pronunciation such as "any vowel" or "any mid vowel" without resorting to a fixed number of alternate phonetic pronunciations. This approach reduces the search space by simplifying the pronunciation models and provides generalization to pronunciations that the creator of the lexicon may not have considered. Finally, we note the possibility of using distinctive phonetic features in a chained HMM, which may allow the features to overlap naturally and provide better coarticulatory modeling.

The topic of distinctive phonetic features can be split into two issues: what features will be used, and how these features will be combined into a phoneme-level representation. The number of possible distinctive features is quite large; a published review of distinctive features revealed about 40 different feature combinations used in various research studies (not necessarily providing unique coverage of all phonemes) [38]. As a result, the set of distinctive features used in the research and development of a speech recognizer and the values that these features can acquire will depend on the goals of the research.

5.3.2 Previous Work on Distinctive Phonetic Features

Kirchhoff used a set of five features, where each feature had from three to ten values [77]. Each feature was learned with a separate neural network, and the distinctive-feature network outputs were combined with the use of another network that mapped distinctive feature values to phonemes. Kirchhoff found that an HMM/ANN system trained in this way had word-level performance comparable to a baseline HMM/ANN system. Combination of the baseline and distinctive-feature systems by multiplication of the phoneme-level neural-network outputs resulted in significantly better word-level performance across a

range of noise conditions.

Koreman, Andreeva, and Barry [82] used Kohonen networks (neural networks that are capable of unsupervised learning) to classify standard cepstral features into 14 distinctive-feature values using three classes (Place, Manner, and Voicing). The distinctive-feature outputs from the Kohonen networks were used as input to a standard HMM system for recognition of consonants in English, German, Italian, and Dutch. They found that this system performed much better than a baseline HMM system on infrequently-occurring consonants, especially language-specific consonants. This supports our motivation of using distinctive phonetic features for multi-language system development.

In work related to the motor theory, Deng and Sun [42, 41] created an automatic speech-recognition system that recognizes five articulatory features (lips, tongue blade, tongue dorsum, velum, and larynx) as an intermediate stage between the acoustic signal and its phonetic representation. Each feature has a range of values; five values for the lip positions, seven for the tongue blade, twenty for the tongue dorsum, two for the velum, and three for the larynx. In this system, changes in the feature values are not required to be synchronized with the phonetic boundaries, allowing a more flexible modeling of coarticulation. The combination of distinctive features is obtained by using separate HMM states for the different possible feature combinations; the total number of required states is reduced by using context-independent categories. Evaluation of this system on the TIMIT database resulted in performance comparable to a context-dependent HMM system, and significantly better performance than a context-independent baseline HMM.

In similar work, Erler and Freeman [44] used a feature set with seven distinctive features, where each feature consisted of a set of between two and six values. These features were derived from the work of Browman and Goldstein on articulatory speech synthesis. This larger feature set resulted in over 7000 states in an ergodic (fully-connected) HMM. Recognition through the HMM states was performed using analysis-by-synthesis instead of the standard Viterbi search, "in order to determine the intended articulations" rather than the realized articulations. In the analysis-by-synthesis method, candidate utterances are converted to articulatory targets, and rules are applied to find the possible paths through the model represented by the candidate targets. The likelihoods of these paths

are computed given the input speech observations, and the path with the highest likelihood is selected. They obtained phoneme-level accuracy of 79.6% on a large-vocabulary isolated word task, which is comparable to the baseline result of 79.9%.

Hübener and Carson-Berndsen [69] trained GMM classifiers on 24 binary-valued distinctive features; the features were motivated by the work of Hess and modified according to performance of the classifiers. They obtained frame-level phonetic accuracy from 77% to 98% on a single speaker, depending on the type of features used to specify the phoneme. They then used the distinctive-feature outputs in two ways: (a) to train an HMM system, which resulted in phoneme-level performance comparable to a baseline system, and (b) as input to a "phonological event parser" that uses distinctive features to construct syllable and phonological hypotheses. This parser is able to specify top-down constraints, allowing only permissible syllable and phonetic structures. Using this parser, they obtained a phoneme recognition rate of 73%.

Schmidbauer [131] used the distinctive features of Manner, Place, and Height, with seven values for Manner, seven Place features, and five Height values. These features were then used as input to an HMM system for phonetic classification. Schmidbauer used formant values, energy, zero crossings, and a "voice-bar" feature instead of cepstral-domain features in order to classify the distinctive features. Evaluation of phoneme-level performance on a set of three speakers resulted in a 10% reduction in error compared to a baseline (context-independent) HMM system trained with mel-frequency cepstral features.

As noted in Section 4.2, Dalsgaard and Andersen [35, 37] used distinctive features to phonetically align speech in an HMM framework. Dalsgaard and Andersen used a self-organizing neural network to estimate the probabilities of distinctive phonetic features. These features were subject to principle component analysis to determine the most relevant features for phonetic classification. The principle components were used as input to a standard HMM system. Distinctive features were used so that the system would be easily applicable to new languages. When evaluated on English speech from the EUROM0 corpus using 15 principle components, agreement was 77.1% within 20 msec.

5.3.3 Proposed Approach for Distinctive Phonetic Features

In our approach, we use three distinctive features (Manner, Place, and Height [1]), where each feature has a set containing six to eleven values. The basis for our feature set comes from Ladefoged [85], although we have modified it slightly to ensure coverage of 43 English phonemes using an economical number of features. For steady-state phonetic categories, we use a combination of context-dependent and context-independent categories, depending on the type of feature. Training and recognition is done with a neural-network classifier using standard cepstral-domain features as well as the acoustic-level features described in Section 5.1 and Chapter 6. The distinctive-feature information that is output from the neural network is combined using Massaro's Fuzzy-Logic Model of Perception (FLMP) [100] to obtain a context-dependent phoneme-level representation prior to the Viterbi search. Both phonetic steady-state and transition probabilities are obtained in this way. This approach to using distinctive features will be discussed in more detail in Chapter 7.

The contribution of this thesis to the area of phonetic modeling in HMM systems is in the combination of distinctive phonetic features using perceptually-motivated rules to arrive at context-dependent within-phoneme probabilities as well as phonetic-transition probabilities. Previous approaches to integrating distinctive features have either used a "higher-level" GMM or neural-network classifier to classify the distinctive features into phonemes, or the states in the HMM have been specified as a combination of distinctive features. Because one of our motivations is to increase the amount of training data per node in the network, passing the distinctive feature values to another phoneme-specific classifier would defeat this purpose. The FLMP provides a perceptually-motivated approach that does not require training, but still maps from distinctive features to phonemes. Although the FLMP has been illustrated theoretically in terms of distinctive features [111], a complete speech recognition or alignment system using FLMP for combination of distinctive features has not been reported in the literature.

[1] We will follow Ladefoged's example [85] and use capital letters when naming a feature, and enclose the values that the feature may have in square brackets.

5.4 System Overview

The proposed alignment system can now be described in terms of its component parts. This system is illustrated in Figures 5.1 and 5.2. Figure 5.1 shows the inputs to one of the neural networks, with the standard PLP feature set as well as the other acoustic-level features. The number of outputs from all networks is much smaller than if context-dependent phonemes were being classified. Figure 5.2 shows how each of the networks from Figure 5.1 are connected; there are three distinctive-feature networks for the within-phoneme classification, with one network each for estimating Manner, Place, and Height. These three networks are combined using FLMP to arrive at a phonetic-level representation. In addition, there are three distinctive-feature networks for phonetic transition estimation, with one network each for Manner, Place, and Height. These transition networks are also combined using FLMP. Finally, the within-phoneme probability outputs are used to estimate the observation probabilities, and the phonetic-transition outputs are used to estimate the state transition probabilities, in an HMM framework. This combination of networks is illustrated in Figure 5.2 with the within-phoneme classifier on top, the transition classifier on the bottom, and the synchronous traversal of the within-phoneme categories and transition probabilities indicated by double vertical lines.

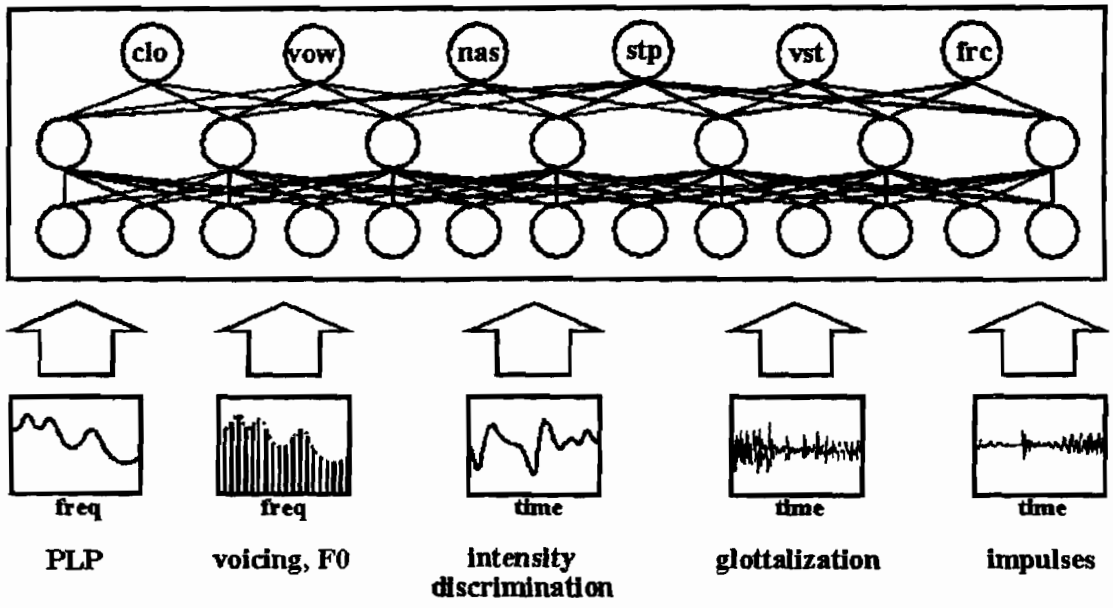

Figure 5.1: Features and neural network for proposed method.

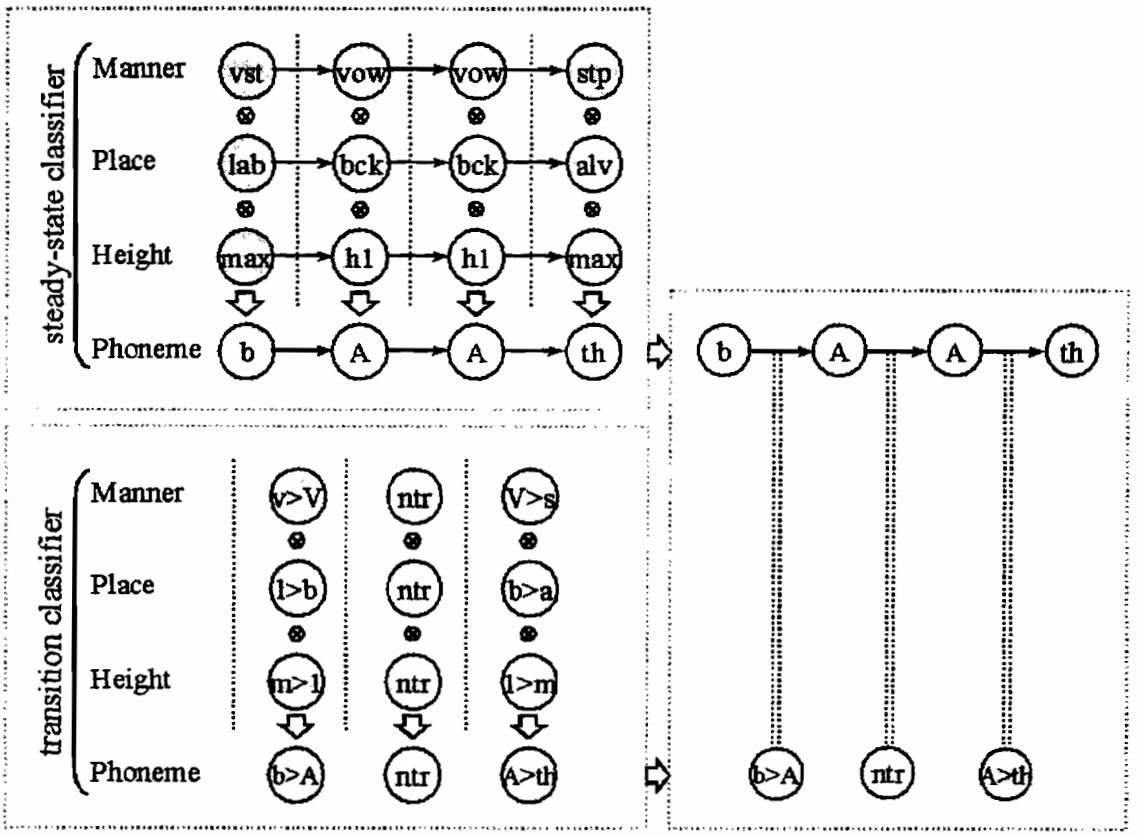

Figure 5.2: Combining the distinctive feature outputs using within-phoneme and phonetic transition networks in the proposed method.

Chapter 6

Acoustic-Level Features

In this section, the acoustic-level features that are used as input to the neural network classifier will be described in detail. As mentioned in Chapter 5, a set of five features was selected for consideration; these features are used in addition to the standard PLP features that represent the spectral-domain information with a 16-msec window.

6.1 Intensity Discrimination

Intensity discrimination is motivated by perceptual studies on the smallest detectable change in intensity, conducted by several psychologists and summarized by Moore [102]. In the psychological studies, two-alternative forced-choice experiments were conducted, and subjects were asked to indicate which of the two stimuli contained the signal with an increased intensity. As Moore reports, despite variations in the methods and stimuli, a general pattern of intensity discrimination is clear. This pattern can be modeled as follows:

$$\Delta L = 10 \log(\frac{I + \Delta I}{I}) \tag{6.1}$$

where ΔL is a measure related to the perceived change in intensity, I is the intensity of the signal, and ΔI is the change in intensity. Intensity, as defined by Moore, is the sound power transmitted through a given area in a sound field, although it can also be used to describe "any quantity relating to the amount of sound, such as power or energy" [103]. A fixed threshold for a change in perceived intensity can be determined; if the absolute value of ΔL is below this threshold, then the intensity change is not detected. Typical thresholds for detection are between 0.5 and 1 dB.

When classifying phonetic transitions in automatic phonetic alignment, we believe it will be useful to have some measure of change in intensity that is related to perception; a small change in absolute energy during a (loud) vowel is of far less importance than a small change in absolute energy during a silent region (which may indicate the beginning of a plosive phoneme). As a result, we apply the formula given by Moore directly, with intensity measured by the energy of the signal, and the window sizes for computing I and ΔI dependent on whether we are interested in long-term or brief changes in the signal. For a general measure of phonetically-relevant changes in the signal, we use a window size of 250 msec for I and a window size of 40 msec for ΔI. The window length for I was chosen to correspond to roughly the duration of one syllable, and the window length for ΔI corresponds to the minimum duration of a speech segment required for assigning phonetic quality and "the interval in which acoustic stimulation begins to assume an independent identity," as reported by Greenberg [59]. As can be seen in Figure 6.1, this measure of intensity discrimination provides a reasonable indication of the onsets and offsets of major phonetic events, as local maxima and minima in the intensity-discrimination measurement correspond to major phonetic changes. Inspired by the work of van Santen and Sproat on multi-band phonetic alignment [141], the work of Sharma and Hermansky on multi-band speech recognition [138], and the Fletcher-Allen model of speech perception [1], we compute intensity discrimination not only for the entire frequency band, but also for seven bark-scale frequency bands, each with a width of one bark.

This measure of intensity discrimination may be useful not only as a feature for detecting phoneme-level changes in the signal, but also for detecting other events in which changes in intensity are a factor. This will be elaborated on in the following sections on voicing, glottalization, and burst detection.

6.2 Voicing

6.2.1 Previous Work on Voicing Determination

A significant amount of research has been done on reliable determination of voicing; a brief review of 20 conference proceedings from 1981 through 1998 yielded 15 papers on

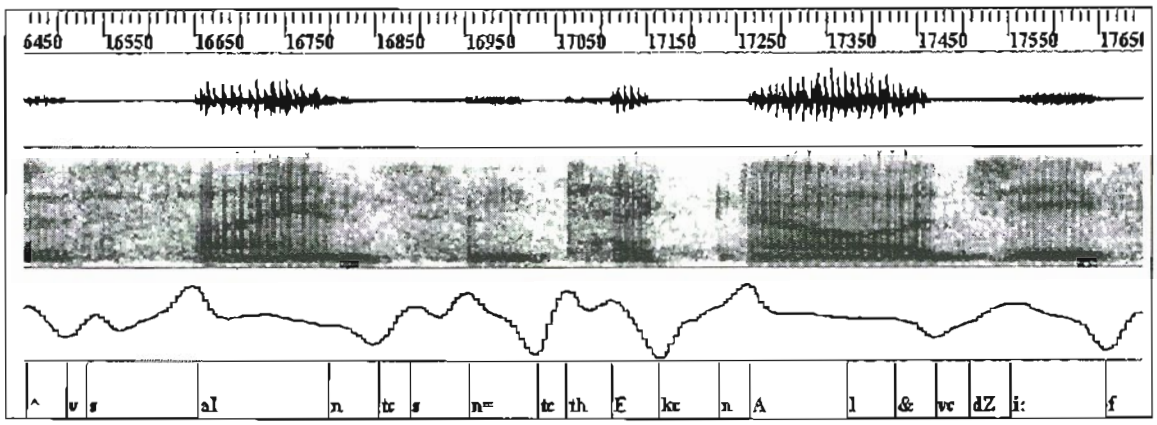

Figure 6.1: Intensity discrimination for continuous telephone-band speech (for the utterance "of science and technology f[or]"). Each panel shows the following: (a) time marks, (b) waveform, (c) spectrogram, (d) intensity discrimination of the entire frequency range, and (e) manual phonetic labeling of the utterance.

the topic (2 in 1998 alone); we were able to find an additional 13 journal articles. The number of papers on this topic and its continued presence as a subject of research is an indicator not only of its importance, but also of its difficulty. Determination of voicing is particularly difficult on telephone-band speech, where the lower 200 or 300 Hz region of the signal (where the fundamental frequency is usually found) has been severely attenuated.

Methods of voicing determination can usually be grouped into one of three general categories: frequency-domain signal analysis methods, methods that use filtering of the signal followed by autocorrelation, or statistical pattern classification methods using features from (both) the frequency and time domains. We will review three of the most common methods (one from each category) and a method that is similar in several respects to our proposed method.

One of the oldest techniques for voicing determination is based on the *cepstrum*, which is simply the spectrum of the log spectrum of a signal. This method, proposed by Noll in 1967 [110], is motivated by the fact that the log spectrum of a signal containing voiced speech will have harmonics at multiples of the fundamental frequency, given an analysis

window longer than one pitch period. These harmonics can then be viewed as a periodic signal that can be subject to further frequency analysis; computation of the spectrum of this harmonic signal will result in a sharp peak with a "quefrency" location (X-axis value) directly proportional to the period of the original speech waveform. A voicing decision can then be made based on the strength of this "cepstral" peak, using a fixed threshold as a decision boundary. Noll proposed weighting the cepstral values more heavily toward higher quefrencies, and weighting the voicing decision based on the values of the surrounding cepstral peaks.

Another approach is called the Simplified Inverse Filter Tracking (SIFT) method, which was proposed by Markel in 1972 [99]; variants on this approach are quite common in the literature. The SIFT method, as defined by Markel, works as follows:

1. The speech is low-pass filtered to 800 Hz and downsampled to 2000 Hz,
2. A 32-msec Hamming window is applied to each analysis frame,
3. LPC analysis of order 5 is applied to the windowed speech,
4. Inverse filtering of the down-sampled speech is performed, using the filter coefficients from LPC analysis,
5. Autocorrelation of the inverse-filtered signal is done,
6. The largest autocorrelation peak is found,
7. Interpolation is done on the region containing the peak, to estimate the height and location of the peak with reduced quantization error,
8. A voicing decision is made, based on the height of the interpolated peak.

According to Markel, the low-pass filtering is done in order to reduce the computational load. The LPC analysis and subsequent inverse filtering remove the formant structure from the signal, resulting in a signal that (in theory) contains peaks directly related to the fundamental period. These peaks occur when there is an abrupt change in the signal that is not well predicted by low-order LPC analysis; such events are common at the instant

of glottal closure. Autocorrelation is used to detect periodicity in the inverse-filtered waveform, and interpolation is done to address quantization error. The voicing decision is made by comparing the autocorrelation peak to a fixed threshold, as well as to previous voicing-decision values.

A third method of voicing determination is to treat it as a problem for statistical pattern recognition. One of the earliest of these approaches was developed by Atal and Rabiner in 1976 [4]. In this method, a GMM classifier is used, with five input features that are known to be correlated with voicing. These five features are computed for each frame with a window size of 10 msec; the features include the energy of the signal, the number of zero-crossings in the waveform, the spectral slope (as measured in dB/octave, which is the first coefficient from LPC analysis), the second coefficient from LPC analysis, and the normalized 12th-order LPC prediction error. Using these features as input to the GMM, they reported 98% frame-level accuracy on a test set of two speakers, each reading one sentence recorded in an anechoic chamber. A similar approach was reported recently by Suh et al. [135], who used six features as input to a recurrent neural network. These features were the energy in the frame; a modified zero crossing rate, called the level crossing rate; the derivative of the level crossing rate; the normalized energy in the range from 180 to 1000 Hz; the normalized energy in the range from 1000 to 2300 Hz; and the normalized energy in the range from 180 to 4000 Hz. Normalization of the energy bands was done by subtraction of the energy in the range from 4000 to 8000 Hz. They reported 92.5% accuracy on a test set of 44 spontaneous sentences read by 17 speakers. They then reported that integration of the voicing feature into a 5000-word recognizer resulted in a 9% reduction in error. Using a similar technique (but with different features and using a feed-forward neural network), Hosom [65] reported 93.4% voicing classification accuracy on a test set of 500 telephone-speech digit utterances; incorporation of this voicing information into a digit-recognition system also yielded a 9% reduction in error.

A method of pitch extraction proposed by Fujisaki and Tanabe [49, 50] is, in its implementation, similar in several respects to our proposed method for voicing determination. The motivation for their method was in observing that the spectral envelope of a signal can be effectively removed by computing spectra with two window lengths. In their method

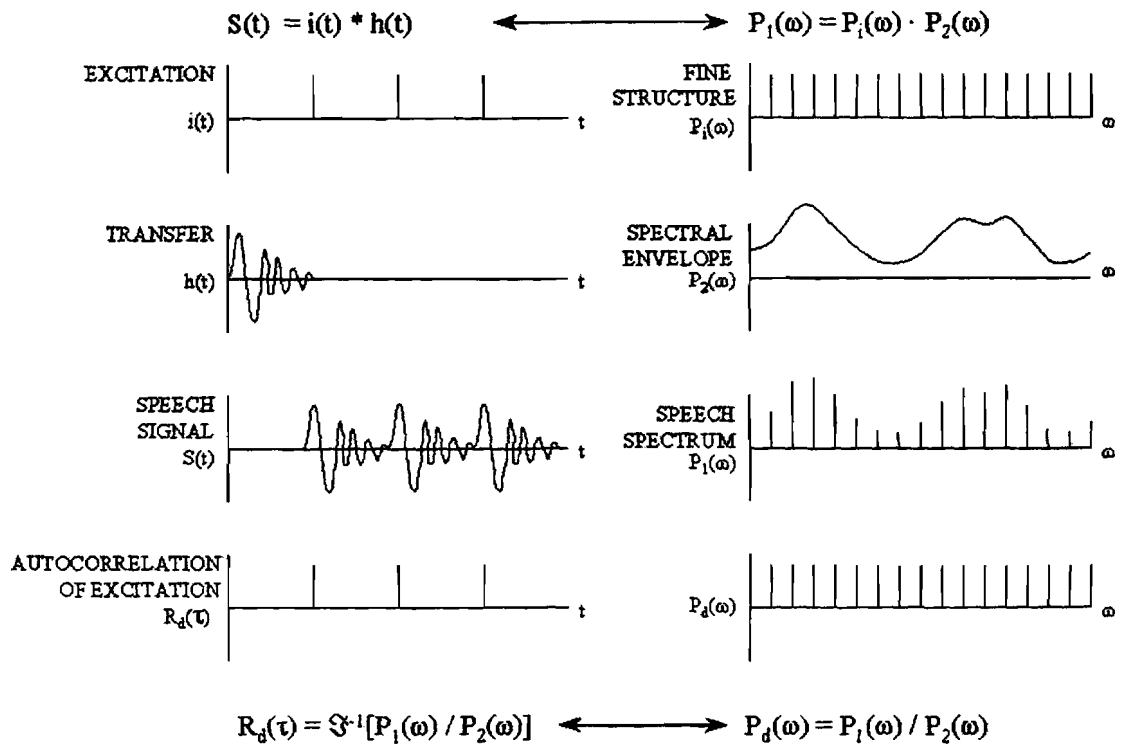

Figure 6.2: Pitch extraction method proposed by Fujisaki and Tanabe. (Figure from Fujisaki and Tanabe, 1972).

(illustrated in Figure 6.2), the power spectrum of a signal with a window approximately equal to one pitch period is computed, resulting in the spectral envelope of the signal without any harmonics. This spectrum is called $P_2(\omega)$. Then, the power spectrum of the same signal, but with an analysis window of several pitch periods, is computed. This spectrum, called $P_1(\omega)$, has the same envelope as $P_2(\omega)$ but includes harmonics at multiples of the fundamental frequency. Then the effect of the vocal tract resonances can be removed by dividing the small-window spectrum by the large-window spectrum, yielding a spectrally-flat series of harmonics called $P_d(\omega)$. The inverse Fourier transform of $P_d(\omega)$ can be considered the autocorrelation of the excitation source; if the source is periodic, then the autocorrelation will contain a peak that has a time position corresponding to

the fundamental period of the source. As Fujisaki and Tanabe note, the autocorrelation is "exempt from subsidiary peaks due to formants, and yet can serve as a measure of periodicity."

6.2.2 Proposed Method for Voicing Determination

Our proposed voicing-determination method is inspired by viewing a spectrogram with a very short analysis window; a sample waveform and its corresponding narrow-window ("wideband") spectrogram is shown in the second and third panels of Figure 6.3. When viewed in such a way, the pitch pulses in the signal are identifiable by dark regions of energy at regularly-spaced intervals corresponding to the fundamental frequency, as noted by Rabiner [118]. This voicing information is present at frequencies greater than 300 Hz, as well as in speech with severely attenuated higher frequencies such as nasal sounds. Our proposed method is motivated by applying intensity discrimination with a very short analysis window to identify periodic energy changes in the speech signal.

In the proposed method, we first band-pass filter the speech from 160 to 700 Hz, thereby removing strong low-frequency noise found in breathy sounds and some plosives, and focusing analysis on the region of the first formant. Then we compute the intensity discrimination of the filtered speech using short-duration analysis windows: 45 msec for the baseline intensity (I) in the denominator, and 4 msec for the change in intensity (ΔI) in the numerator. The window size of 45 msec corresponds to a duration of at least two pitch periods, as we want to ensure that this reference intensity level includes any F0-related impulses. The window size of ΔI must be short enough to detect pitch-related energy changes within even short pitch periods. The resulting intensity discrimination yields a series of regularly-spaced pulses for voiced speech (corresponding to the periodic energy changes resulting from opening and closure of the vocal folds), and irregular pulses for unvoiced speech (corresponding to aperiodic energy fluctuations in the region of the first formant). An example pulse train is illustrated in the fifth panel of Figure 6.3. Autocorrelation is applied to this pulse train, and thresholds (based on the autocorrelation peaks, relative energy levels, and durations of the autocorrelation peaks) are used to determine the voiced regions. The peaks in the autocorrelation result are enhanced by

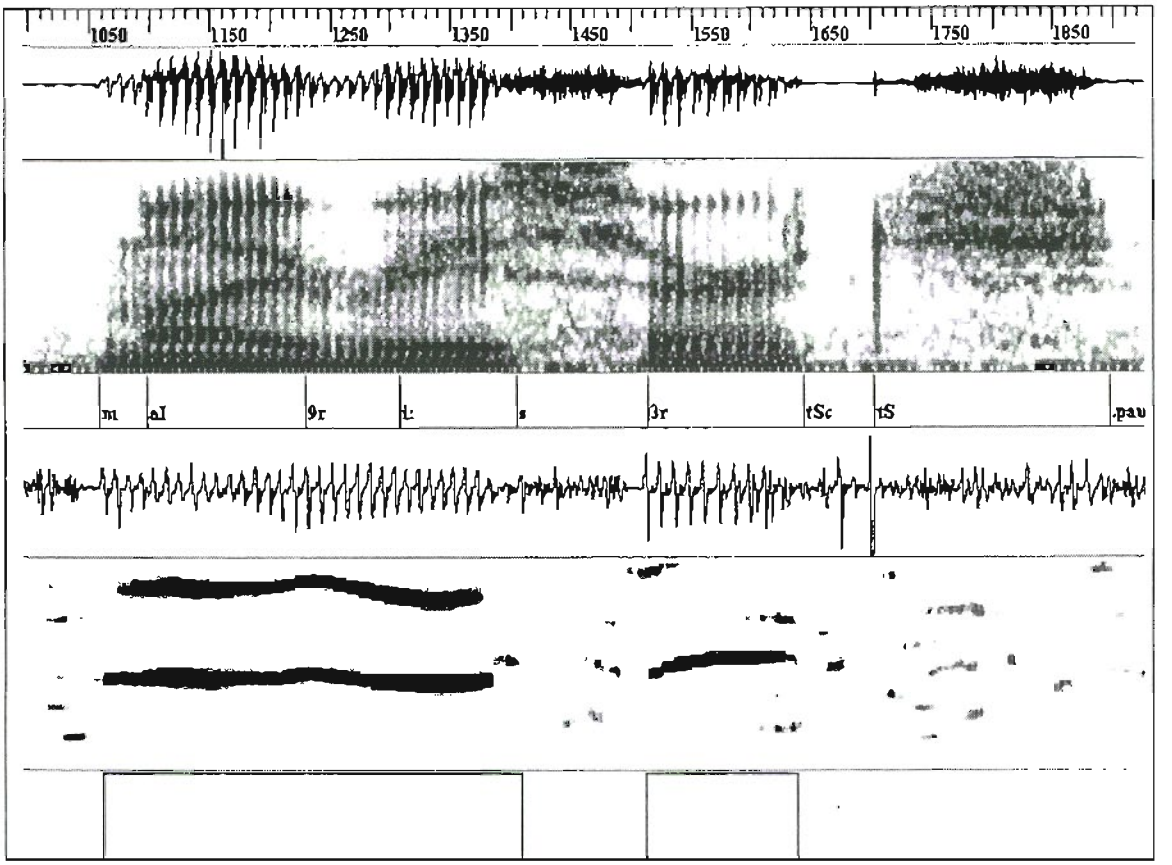

Figure 6.3: Illustration of voicing computed by proposed method for the utterance "my research." Panel (a) shows time marks in msec, (b) shows the waveform, (c) shows the broadband spectrogram, (d) shows the phonetic transcription, (e) shows the computed pulses train, (f) shows the enhanced autocorrelation, and (g) shows the binary voicing decision from the proposed method. A frame rate of 0.5 msec is used for visual clarity.

performing two autocorrelation computations for each frame: one forward in time from the current frame, and the other backward in time from the current frame. Making the (somewhat dubious) assumption that these two autocorrelation results are independent because they occur at different times in the signal, we combine the two results at each frame to obtain one "enhanced" autocorrelation value per frame. This technique has the effect of accentuating the peaks and valleys of the autocorrelation result, and in informal tests provides better voicing-determination results than without the enhancement. The proposed voicing-determination method uses 17 parameters, which were first set to initial

values based on general knowledge of their effects, and then refined by iteratively changing each parameter and evaluating its effect on a development set of about 60 utterances of hand-labeled speech.

Voicing determination is then defined by our model as the ability to detect periodic intensity changes in the first-formant region of the signal. This definition is built upon our knowledge of the voicing source, which specifies periodic changes in intensity affecting all frequency bands, and our knowledge of the sounds of voiced speech, which have a strong resonant energy in the lower frequency region (160 to 700 Hz) due to the first formant.

An advantage of the proposed method over SIFT-based methods is that LPC analysis of the signal is not required in the proposed method; the error peaks resulting from LPC analysis in SIFT are not guaranteed to correspond directly to glottal pulses. Specifically, the shape of the LPC residual may vary depending on the analysis order, the abruptness of glottal closure, and the amount of noise in the signal. An advantage of the proposed method over cepstral-domain methods is that our method requires only a few harmonics for accurate voicing determination; cepstral-based methods require harmonics throughout the spectrum. (Harmonics may be lacking in the higher frequencies due to noise or to spectral zeros related to the production of lateral or nasal sounds.)

In the statistical pattern-classification approach, the correlations between the input features and voicing result is not always a cause-and-effect relationship. For example, (unvoiced) breath noise may have a spectral slope characteristic of voiced speech, (unvoiced) whispered speech may have the formant structure of (voiced) vowels, front vowels in additive noise may have a very high zero-crossing rate normally associated with unvoiced sounds, and unvoiced bursts may have large energy usually associated with voiced sounds. As a result, the overall correlation between the features and the voicing property can be low under several conditions. The proposed method bases its decision on features directly related to the consequences of voicing, and so the features therefore have a higher correlation with voicing under various circumstances than the features used in the pattern-classification approach.

Finally, the proposed technique is different in several respects from the one proposed by Fujisaki and Tanabe; in their method, the window length of the narrow window should

be about one pitch period, whereas in our case, the narrow window (corresponding to ΔI) should be much shorter than one pitch period. Also, in the Fujisaki and Tanabe approach, the entire spectrum of the signal is computed for both the narrow and wide windows, with the intent being to normalize for the shape of the spectral envelope. If some frequency regions of the signal contain noise with a higher energy level than the harmonics, then the energy level of this noise at those frequencies will be made equal to the energy level of the harmonics at other frequencies. This strong noise level at different regions of the spectrum may yield spurious peaks in the subsequent inverse Fourier transform. In the proposed method, the relative energy of the signal at one frequency band is computed for the narrow and wide windows, with the intent being to capture the time-synchronous relative changes in the energy due to the glottal-source excitation; noise at frequencies beyond the 200 to 700 Hz range does not affect the proposed method.

Currently, we use only one frequency band for voicing determination, because the harmonics of voiced speech do not always occur at the same instant in time. (The vertical bands of energy corresponding to a glottal pulse may shift slightly as a function of frequency.) As noted by Moore [104], it is likely that humans use multiple frequency bands when determining pitch; the use of multiple frequency bands may apply to voicing determination as well. The use of multiple frequency bands would be a simple extension of our proposed method; in the same way that a single band is now used, voicing determination for multiple bands could be computed, and the autocorrelation results then combined using a technique such as voting. The reason why such an extension has not been pursued as part of this thesis is the time required for autocorrelation. Given our current processing power (200 MHz Pentium Pro) and the requirement that our final phonetic-alignment system operate significantly faster than manual alignment, the use of multiple frequency bands was considered prohibitive.

Given results from voicing determination, the voice-onset time (VOT) can be defined as the time from the current frame until the closest change from unvoiced to voiced speech. A limit of 150 msec is imposed on the search for the closest change in voicing, as changes beyond 150 msec are not likely to be related to the phoneme at the current time frame. For this thesis, voicing information is composed of both a binary measure of the presence

of voicing, and a measurement of VOT.

6.2.3 Results of Voicing Determination

One weakness of most papers on voicing determination is that analysis is typically done on a fairly small test set, usually from a small number of speakers and recorded over a single channel. Very few papers evaluate their proposed method on a publicly-available corpus or over different channel conditions. To evaluate the relative merit of our proposed algorithm, we implemented the SIFT algorithm as described by Markel, and modified it to use our proposed method's technique of merging short-duration frames (in which voiced speech is required to have a minimum duration of 30 msec, and unvoiced speech is required to have a minimum duration of 15 msec). We then evaluated the original SIFT method, the modified SIFT method, and our proposed method on the test portions of the TIMIT [52], OGI Stories, OGI Numbers, OGI Portland Cellular, and OGI Kids' corpora [27, 132]. In order to focus on the regions of speech in which voicing is easily determined, and in order to avoid having to hand-label these corpora according to their voicing characteristics, we mapped each phoneme to a type of voicing. Vowels, nasals, retroflex sounds, liquids, and glides were mapped to voiced speech; unvoiced fricatives, unvoiced plosives, unvoiced closures, and silence were mapped to unvoiced speech; all other phonemes (voiced fricatives, voiced bursts, and voiced closures) were not evaluated as their voicing properties may vary depending on context and speaker style. In addition, the 10 msec before and after each phonetic boundary was not evaluated, because humans disagree on phonetic boundary placement about 20% of the time with a 10-msec interval. Evaluation was done on a frame-by-frame basis.

As can be seen in Figure 6.4, the proposed method is significantly better than either of the SIFT methods, with reductions in error of at least 58%, 39%, 68%, 83%, and 42% for the TIMIT, Stories, Numbers, Portland Cellular, and Kids' corpora, respectively. Note that the SIFT results may not represent state-of-the-art in voicing determination, but they provide a well-known baseline against which other methods can be compared. The accuracy of the pattern-recognition-based voicing determination method developed by Hosom in 1996 and evaluated on the OGI Numbers corpus [65] is included in Figure 6.4;

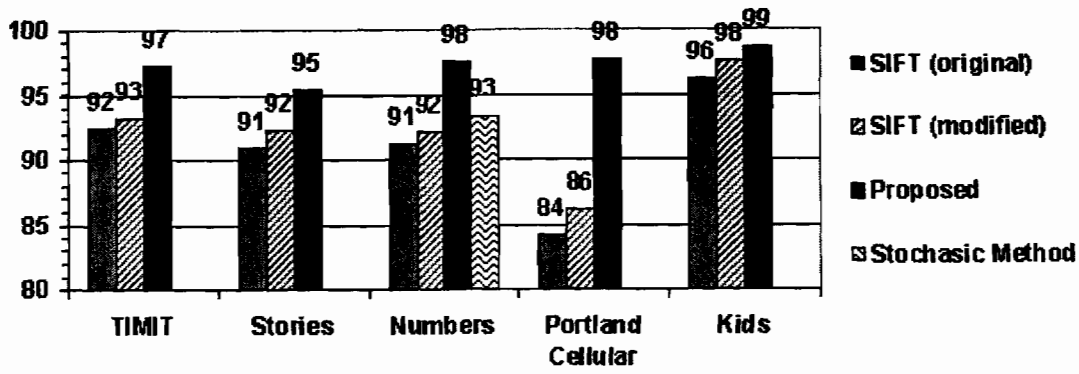

Figure 6.4: Results of voicing determination on various corpora (indicated on the horizontal axis) by the original SIFT method, the modified SIFT method, and the proposed method. In addition, results of a recently-developed method similar to Atal and Rabiner's stochastic pattern-classification method is plotted for the Numbers corpus.

this more current method of voicing determination was found to be significantly more accurate than the similar stochastic method developed by Atal and Rabiner [4]. It can be seen that the proposed method still has a 62% reduction in error over this more current method on that corpus. It is also interesting to note that the potentially high fundamental frequency and first-formant values of the children's speech did not adversely affect the performance of either the proposed method or the SIFT method.

It should also be noted that the proposed method performs competitively on cellular-band speech, even though cellular speech was not used at all in the development or refinement of parameters. This indicates that the LPC residual is not a reliable indicator of glottal excitation for cellular speech, whereas the intensity discrimination measurement used in the proposed method is not adversely affected by typical channel conditions. The high accuracy for the Kids corpus may be due to several factors: the children were prompted for single words, which may have resulted in better-than-average enunciation, and because only single words were uttered, glottalization may not have been as prevalent as in a continuous-speech corpus such as TIMIT, Stories, or Portland Cellular.

6.3 Fundamental Frequency

6.3.1 Previous Work on Fundamental Frequency

Development of a robust method of extracting fundamental frequency is still an area of research, despite numerous papers and journal articles on the subject. Extraction of the fundamental frequency, or F0, is related to voicing determination, in that there is no fundamental frequency when the speech is unvoiced, and most or all voiced speech has a fundamental frequency (depending on whether or not one considers glottalized speech to be voiced speech). However, instead of a binary decision about whether or not the vocal folds are vibrating, the rate of vibration of the vocal folds must be estimated.

Methods for determining F0 can be grouped into the same three areas as voicing determination, and many of the methods used for voicing determination require only slight modification to extract F0. For example, in the same way that the strength of a cepstral peak can be used to determine voicing, the location of this peak on the X (quefrency) axis can be used to determine F0.

The SIFT method of voicing determination can also be used to estimate F0, because the location of the autocorrelation maximum on the X (samples) axis is an estimate of the fundamental period (in samples). We have modified the SIFT algorithm to extract F0 values with greater accuracy. In the original SIFT algorithm, LPC analysis of order 5 is used. This value produces good results for voicing determination, but for pitch extraction a secondary peak in the autocorrelation result, corresponding to a strong first formant, is often observed. Such a secondary peak can cause the extracted F0 value to be double or triple the correct value. As Markel was concerned with a fast implementation, he was constrained to use only one LPC analysis for voicing and F0 determination. Given current processing power, however, we performed voicing determination using LPC order 5, and on the frames of voiced speech we performed F0 extraction using a second LPC analysis with order 9. This higher order was determined by evaluating F0-extraction results of a small development set using orders 5, 7, 9, and 11.

A pattern-classification approach to F0 extraction was developed by Barnard, Cole, Vea, and Alleva [5]. They used a neural network classifier which had as its inputs the

amplitudes of peaks in a low-pass filtered waveform and their time differences. This network then classified each input peak as associated with a glottal pulse or a formant resonance. The pitch could then be computed from the peaks identified as glottal pulses. They reported accuracy of 97.5% on a 20-speaker subset of the TIMIT corpus, classifying each peak in voiced speech as glottal-pulse-related or not, and comparing these outputs with manual determination of which peaks were related to glottal pulses.

In a method developed by Harris and Nelson [62], phase tracking of the pulses in voiced speech is accomplished by correlating the speech signal with a time-varying filter that is matched to previous waveform pulses. An initial filter is created from the first identifiable pulse in the waveform, with the pulse centered within the filter. Future pulses are determined by an iterative process of matching the current filter with the signal and then updating the filter. The current filter is matched with the signal by computing the minimum distance:

$$d(f, g_\tau) = min \;\; \alpha, \beta, \tau \{ \int |f(t) - (\alpha g_\tau(t) + \beta)|^2 dt \} \tag{6.2}$$

where $f(t)$ is the current filter, $g(t)$ is the waveform, τ is a variable timing offset for the waveform, and α and β are arbitrary scalars for which $d(f, g_\tau)$ is minimum. Harris and Nelson give an efficient method of determining this minimum distance, in which α and β are computed directly, and τ is determined by iterating over a set of values that covers the expected range of fundamental periods. The value of τ that corresponds to the minimum distance indicates the position of the pulse, and can be used to compute the fundamental frequency. Filter updating is done by applying a non-uniform weighted average of the most recent filter with the most recently aligned pulse. (The weighted average is maximum at the point of maximum energy of the pulse.) An extension to this procedure accentuates the periodicity of the signal by minimizing points in the waveform with a high variance. This method of F0 extraction, which we will call the Harris and Nelson method, was implemented in the CSLU Toolkit by Johan Schalkwyk, and serves as a baseline method for comparison, along with the SIFT method implemented for this thesis.

6.3.2 Proposed Method of Fundamental Frequency Extraction

In the same way that the SIFT or cepstral methods of voicing determination can be extended to F0 extraction, we have extended our proposed voicing method to return estimates of the fundamental frequency. As in the SIFT method, voicing is determined based on autocorrelation results, and the fundamental frequency is computed from the index number of the autocorrelation peak:

$$F0 = F_s/k \tag{6.3}$$

where F_s is the sampling frequency and k is the index of the autocorrelation result corresponding to the autocorrelation peak. In the SIFT method, Markel considered that the 2000-Hz sampling rate necessitated interpolation of the autocorrelation results to determine the peak location with less quantization error. In our method, we use a sampling rate of 8000 Hz to compute the pulse train, which is reduced during autocorrelation to 4000 Hz to improve processing speed. We then determine the F0-related autocorrelation peak from the highest autocorrelation index. (Unless the pitch is extremely low, there will often be more than one autocorrelation peak, and the higher peaks are required to be at multiples of the index of a primary peak.) The use of these higher autocorrelation peaks to determine F0, with an effective sampling rate of 4000 Hz, results in a quantization step between 0.44 Hz and 1.77 Hz. As a result, we do not use interpolation in determining the F0 value.

A post-processing step is also implemented in our method, in which sudden doubling or tripling of the F0 value is checked for. If such an event happens, an FFT of the speech signal is computed, and a check is made to locate peaks below the harmonic associated with the larger F0 value. If such peaks are found, then it is assumed that the lower pitch is the correct one, and the higher pitch value is reduced. Also, if the pitch value is halved for a duration of less than 30 msec, it is assumed that the higher pitch value is correct, and the lower values are multiplied by a factor of two. If the voicing determination indicates that a frame is voiced but there is no autocorrelation peak, then the F0 value is interpolated from surrounding peaks. Finally, spurious peaks and dips in the F0 contour are removed. It is arguable that in some cases (such as at the end of a word) the F0 value may in fact

suddenly decrease, and that such post-processing is fixing imaginary errors. However, it can also be argued that such dramatic changes in F0 are, when of short enough duration, not relevant to the phonetic or perceived content. As our interest is in F0 changes that are related to phonetic content, and as we have a separate detector for glottalization (which would presumably detect a sudden decrease in F0 at the end of a word), it is felt that such post-processing is beneficial for the ultimate goal of phonetic alignment.

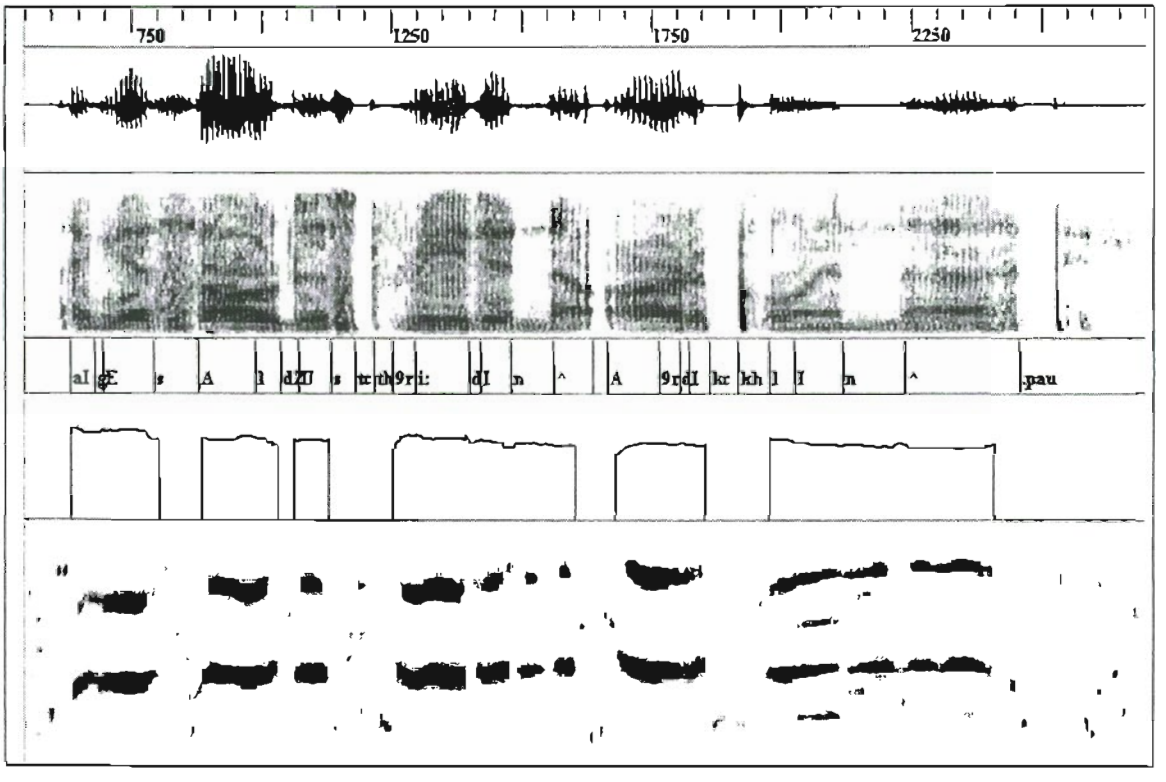

Figure 6.5: Example of F0 extraction for the utterance "I guess I'll just read in a article in a..." (telephone speech). Panel (a) shows time marks in msec, (b) shows the waveform, (c) shows the broadband spectrogram, (d) shows the manual phonetic labeling, (e) shows the pitch values extracted using the proposed method, and (f) shows the enhanced autocorrelation. A frame rate of 0.5 msec is used for visual clarity.

An example of pitch extraction with the proposed method on telephone-band spontaneous speech is given in Figure 6.5. As the voicing decision is made prior to F0 extraction,

several frames have F0 values but no autocorrelation peaks. These regions are all less than 20 msec in length, and the F0 values have been interpolated from surrounding values.

6.3.3 Results of Fundamental Frequency Extraction

A major difficulty in determining the accuracy of a method for extracting F0 is in finding the correct values for the fundamental frequency. In some papers (for example, Barnard et al. [5]), the correct F0 values are determined by visually inspecting the waveform and locating waveform peaks that are related to vocal fold vibration. Because this method of determining F0 values is time-consuming and labor-intensive, evaluations based on such values are usually limited to a small test set. Another solution is to use Electroglottography (EGG) measurements of the instant of glottal closure. The EGG "registers laryngeal behavior indirectly by measuring the change in electrical impedance across the throat during speaking" [98], thereby measuring the glottal vibrations without processing of the acoustic waveform that has been filtered by the vocal tract. The value for F0 is then estimated from the EGG signal by the following equations:

$$F0 = \frac{1}{(p_n - p_{n-1})} \quad \text{if } p_{n+1} - p_n > 2.5(p_n - p_{n-1}) \tag{6.4}$$

$$F0 = \frac{1}{(p_{n+1} - p_n)} \quad \text{if } p_n - p_{n-1} > 2.5(p_{n+1} - p_n) \tag{6.5}$$

$$F0 = \frac{1}{\frac{(p_n - p_{n-1}) + (p_{n+1} - p_n)}{2}} \quad \text{otherwise} \tag{6.6}$$

where F0 is the estimated fundamental frequency, and p_n is the location of the n^{th} pitch mark (in seconds). According to a tutorial written by Krzysztof Marasek of Stuttgart University [98], the EGG is a "very precise and robust carrier of F0 even for moderately pathological voices." As part of the text-to-speech development effort at CSLU, a corpus of 500 files of read utterances from a single speaker with EGG measurements has been created; these data are referred to here as the MWM-EGG corpus. The F0 values were determined for each 5 msec frame using the EGG data.

The F0 values determined from the EGG data of the MWM-EGG corpus were compared with the F0 values determined by the proposed method, the original SIFT method,

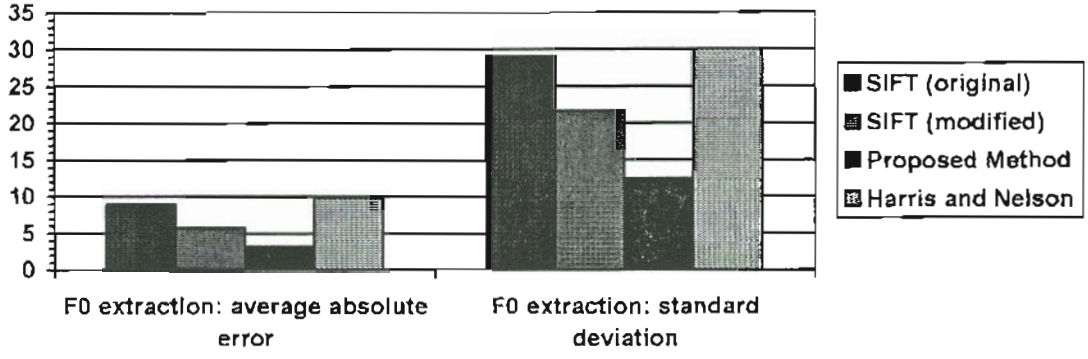

Figure 6.6: Results of F0 extraction for the original SIFT method, modified SIFT method, proposed method, and Harris and Nelson method. Evaluation was performed on the MWM-EGG corpus of read speech, with reference F0 values computed from EGG data. The left group of values shows average absolute error, and the right group shows standard deviation.

the modified SIFT method, and the CSLU Toolkit's implementation of the Harris and Nelson method. At any given frame, a comparison of F0 values was only done if both the EGG result and the measured result had F0 values greater than zero (indicating voiced speech); this was performed to separate the evaluation of voicing determination from F0 accuracy. Results are shown in Figure 6.6. It can be seen that not only is the average absolute error for the proposed method at least 45.3% closer to the EGG data (as compared to the next-best modified-SIFT method), but the standard deviation of the results from the proposed method is 42.7% smaller than the modified SIFT method.

6.4 Glottalization

6.4.1 Previous Work on Glottalization Detection

A literature review of glottalization revealed no automatic methods for determining when speech is glottalized or not, although one paper did determine that glottalization is a characteristic of speech that can be detected by human listeners [70]. We are aware of one

unpublished method of automatic glottalization determination, developed by Cole in the 1970's at Carnegie Mellon University.

In Cole's method, an estimate of the median pitch is used to detect sudden decreases in the fundamental frequency that are characteristic of glottalization. First, the signal is passed through a 1000-Hz low-pass filter to smooth the waveform. Then, the peak-to-peak (PtP) amplitude of the signal is computed at each frame. The window size for determining the PtP amplitude is set to 1.3 times the median fundamental period. Because of this specific window size, pitch periods that are fairly close to or greater than the median value result in large PtP amplitude values. If, however, the F0 value drops below 77% of the median value, then there will be a lack of glottal pulses within an analysis window, and the PtP amplitude will drop. When the analysis window is shifted forward in time so that it includes a glottal pulse, then the PtP amplitude will again increase. As a result, non-glottalized voiced speech has consistently large PtP values, but glottalized speech has dips and peaks in the PtP values. This is illustrated in Figure 6.7, which shows a sample waveform containing normal speech as well as glottalized speech. Glottalization can then be determined based on the presence or absence of spikes in the PtP amplitude signal.

Although this is the only previously-attempted method of automatic glottalization determination that we are aware of, it would also be possible to train a stochastic classifier to identify glottalization given standard PLP features. Such a classifier might be able to account not only for the irregular pulses found in glottalized speech, but also for the resonance characteristics that are present at each glottal pulse. The difficulty in training such a classifier is similar to the problem of using simple energy or waveform values to train a voicing determination classifier: the relationship between the features used as input (time-domain energy or waveform samples) and the desired output classes (periodic signal or aperiodic signal) is quite indirect, and if the network does not learn the correct mapping, then generalization will be poor [5].

6.4.2 Baseline Methods for Glottalization Detection

In order to measure the relative effectiveness of our proposed system (described below in Section 7.4.3), we trained two baseline ANN classifiers that identify glottalization using

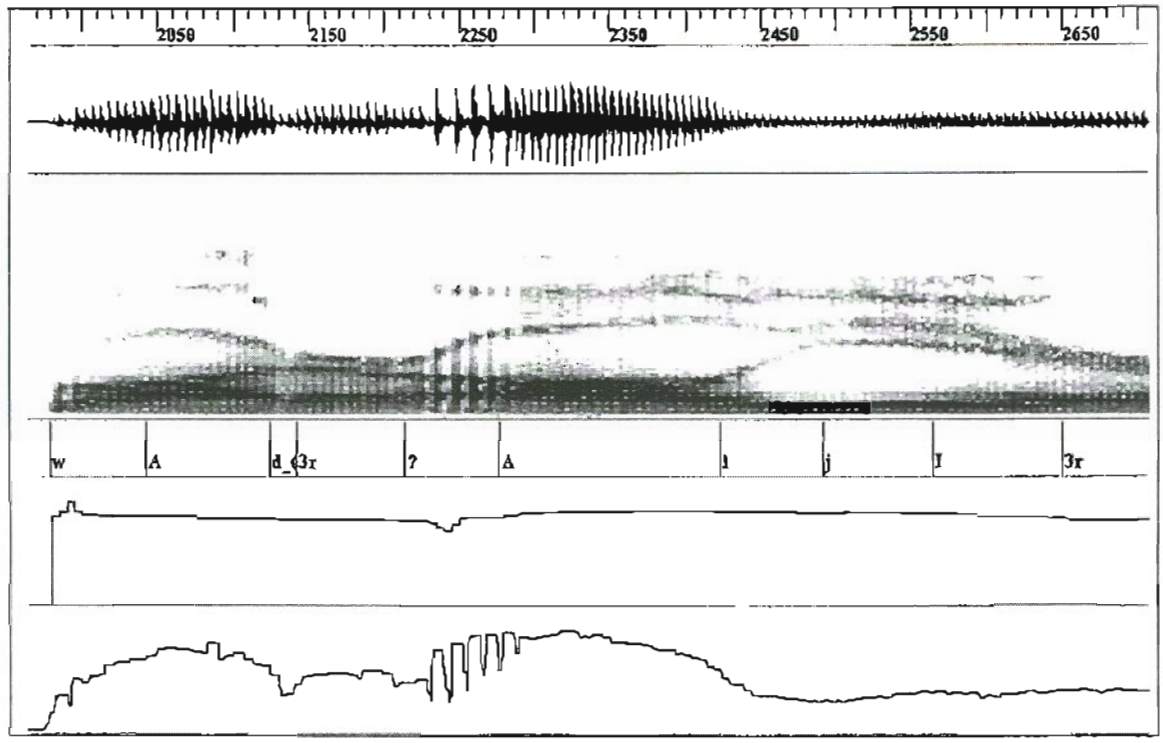

Figure 6.7: Illustration of Cole's method for detecting glottalization for the utterance "water all year." In each panel are (a) time marks, (b) the waveform, (c) the spectrogram, (d) the phonetic labels, (e) the estimated F0 values, and (f) the results of peak-to-peak glottalization detection. The spikes in this bottom output indicate the estimated location of glottal peaks. A frame rate of 1.0 msec is used in this example.

PLP features alone; the two baseline systems differ only in the size of their context window. The first system, referred to as "Baseline A", was trained with the standard PLP feature set (with delta values) and a standard context window of frames at -60, -30, 0, 30, and 60 msec relative to the frame of interest. The second system, referred to as "Baseline B", was trained using the standard PLP feature set (with delta values) and a more dense context window of frames from -60 to +60 msec relative to the frame of interest, at 5 msec intervals (for a total of 25 frames per context window). Both systems have two context-independent output categories, "glottalized" and "non-glottalized." Training was done with approximately 20,000 examples per category. Training, development, and final

evaluation of the two systems were done using the TIMIT, Stories, and Portland Cellular corpora.

6.4.3 Proposed Method for Glottalization Detection

In the proposed method of glottalization determination, we modify the method developed by Cole. In our method, we use intensity discrimination as described in Section 7.1 to locate glottalization peaks, instead of using the PtP waveform amplitude, which can vary according to the microphone characteristics, degree of overall loudness, and type of phoneme. First, a glottalization pulse train is computed using intensity discrimination to identify possible points of glottalization. Then the pulse train values are passed, along with PLP coefficients, to a neural network for final classification. The neural network uses a context window of 25 frames spanning 120 msec. The use of the glottalization pulse train and the PLP coefficients allows classification based on the spectral characteristics of the signal.

We use our proposed method of F0 extraction (described in Section 7.3) to estimate the F0 value for determining window sizes. A window size of 45 msec is used for the reference intensity (I), a length of twice the estimated fundamental period is used for the window size of the intensity change (ΔI), the intensity delta values are computed using a 6-msec window, and a cutoff threshold of 0.06 is used. The reference intensity window size is set to be long enough to include at least two pitch periods, and other values were determined by evaluating a number of possible sets of values. For each set of values, a development-set analysis was done to determine the relative number of insertions and deletions. The resulting receiver-operating characteristics (ROC) curve is shown in Figure 6.8. Based on this figure, we selected three sets of values with which to train and evaluate neural-network based glottalization classification. The results of neural-network-based classification with each set are given in Table 6.1. The set of values that yielded the best classification results (2.0 times the fundamental period for the window size of ΔI, a 6-msec window for computing the delta values, and a threshold of 0.06) was selected as the final set of parameter values.

As in the baseline systems, training was done on 20,000 examples of each of the two

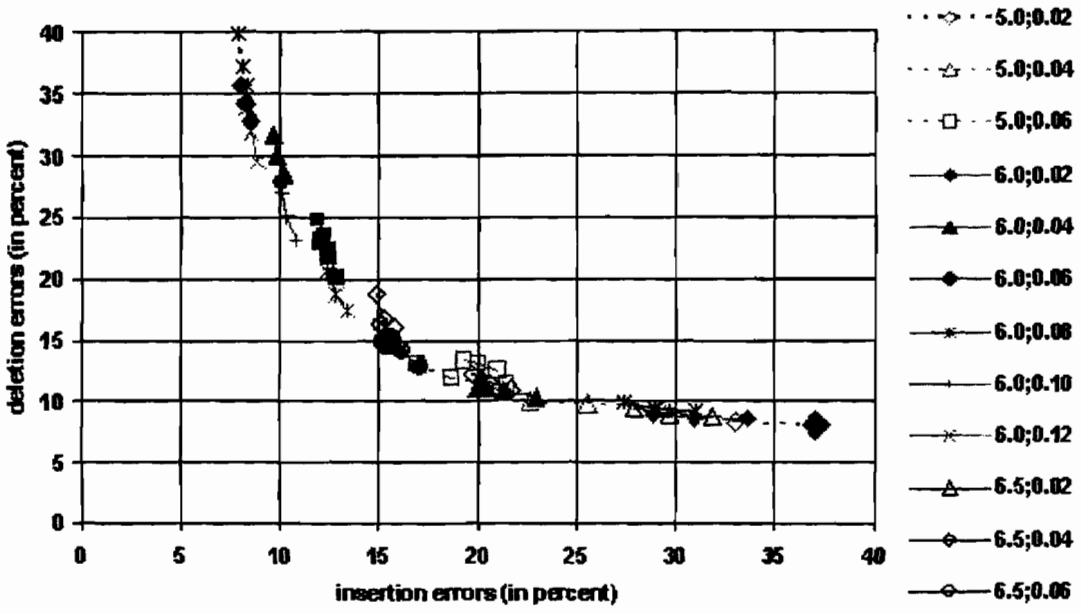

Figure 6.8: Plot of insertion errors (horizontal axis) against deletion errors (vertical axis) for various sets of parameter values, given glottalization pulse train information. Values are connected with lines according to the window size of ΔI for visual clarity; the legend indicates the delta and threshold values for each set of window sizes. The three sets of values chosen for further evaluation using neural-network classification are indicated by the larger points.

categories, and the training, development, and final evaluation were done on the TIMIT, Stories, and Portland Cellular corpora. Output from the proposed method is illustrated in Figure 6.9, including the glottalization pulse train and neural-network results.

6.4.4 Results of Glottalization Detection

Results for the two baseline systems and the proposed method are given in Table 6.2. In these results, the insertion rate is measured according to the number of frames for which glottalization is detected but not present within 20 msec, relative to the total number of non-glottalized frames. The deletion rate is measured according to the number of frames for which there is glottalization but glottalization is not detected within 20 msec, relative to

Table 6.1: Glottalization detection error rates on the TIMIT development set using a neural-network classifier with a glottalization pulse train and PLP features as input. The parameter values used in computing each pulse train are specified in the first three columns.

ΔI window size (msec)	range of delta (msec)	threshold	classification error (%)
2.0	6.0	0.06	11.87
2.0	6.0	0.04	11.92
1.8	2.5	0.02	17.53

Table 6.2: Error rates of glottalization detection for three methods (Baseline A, Baseline B, and Proposed Method) for three test-set corpora.

Corpus	Baseline A error rate (%)	Baseline B error rate (%)	Proposed Method error rate (%)	Relative Reduction in Error (%)	Significance Level (%)
TIMIT	14.16	13.23	12.08	8.69	$p < 0.001$
Stories	20.95	17.43	17.78	-2.01	$p < 0.001$
Portland Cellular	21.27	16.54	16.89	-2.12	$p < 0.001$

the total number of glottalized frames. The reported error result is the sum of the relative insertion rate and relative deletion rate. It can be seen that the performance of Baseline A is always worse than that of Baseline B and the proposed method, but that the relative performance of Baseline B and the proposed method varies according to the evaluation corpus. For the TIMIT corpus, the proposed method has an 8.69% relative reduction in error, but for the Stories and Portland Cellular corpora, the proposed method has relative increases in error of 2.01% and 2.12%, respectively. A statistically significant difference can be claimed between Baseline B and the proposed method for all three corpora, using McNemar's test [55]. Therefore, the use of glottalization peaks in the neural-network classifier does not always yield superior performance, given a sufficiently large context window of PLP features. However, the nearly 9% reduction in error on the TIMIT corpus does, in our opinion, make the use of glottalization peaks worthwhile if extremely high accuracy is desired overall.

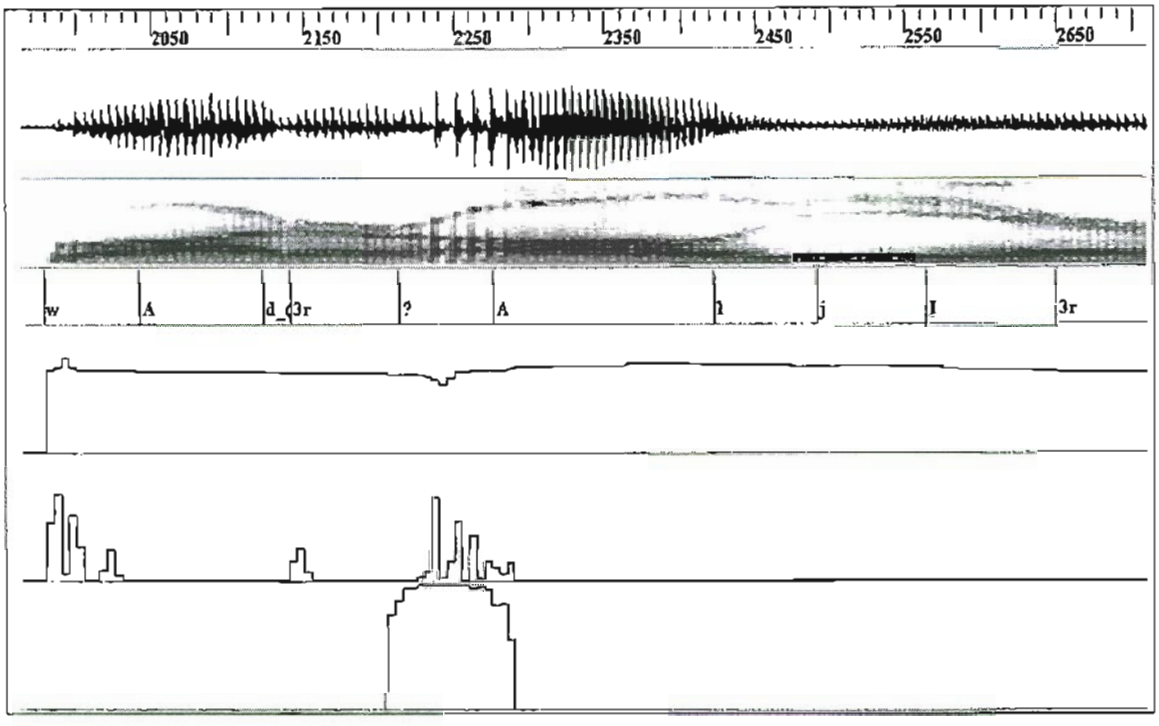

Figure 6.9: Illustration of the proposed method for detecting glottalization for the utterance "water all year." In each panel are (a) time marks, (b) the waveform, (c) the spectrogram, (d) the phonetic labels, (e) the estimated F0 values, (f) the glottalization pulse train (with a 5-msec frame rate), and (g) the results of neural-network classification.

6.5 Impulses

6.5.1 Previous Work on Impulse Detection

Several papers have been published on impulse detection for locating the instant of burst release in plosive phonemes; these methods include neural-network classification [78], HMM classification [109, 108], rule-based methods [97], thresholding of energy derivatives in various frequency bands [91, 107], and the use of support vector machines (SVM) [109, 108]. Here we will describe a derivative-of-energy method (because of its intuitive nature) and the SVM method (because its reported accuracy is the best found in the literature, testing has been done on the commonly-available TIMIT corpus, and a comparison

between the SVM method and other methods has been given).

In the derivative-of-energy method develop by Liu [91], the following procedure is used:

1. A smoothed spectrogram is obtained by first computing the FFT using a 6-msec window at every 1-msec frame, and then smoothing these results with a 20-msec window.

2. The maximum spectral value in each of five frequency bands is determined from the smoothed spectrogram.

3. A measure of change, called rate-of-rise (ROR), is computed by taking differences of the spectral maxima at +25 msec and −25 msec relative to the frame of interest (a 50-msec window), for each frequency band.

4. A cutoff threshold of ±9 dB is applied to each of the five ROR measurements, and peaks beyond this threshold are detected.

5. For each detected peak, the time position of the peak is adjusted using a spectrogram with a 10-msec smoothing window, a ROR computation with a 26-msec window, and a 6 dB cutoff threshold.

6. A voicing determination is made at each boundary based on the ROR measurement of the lowest-frequency band.

7. A burst determination is made based on the ROR measurements of the other frequency bands, with the requirement of a minimum duration of unvoiced speech preceding the burst.

Liu evaluated this method on a corpus of four speakers reading 20 sentences each, with varying amounts of additive noise. As the main focus of this research was on the detection of all acoustic landmarks (including voicing and sonorant/consonant boundaries), evaluation of only the burst-detection component was not reported.

The SVM method, as described in two papers by Niyogi et al. [109, 108], works by classifying a set of binary features ("burst" and "non-burst") using a support vector machine; SVMs are binary classifiers that are considered to provide good generalization

to unseen data. Niyogi et al. implemented two SVMs that are capable of linear and non-linear classification, respectively. The input to the SVM at each 1-msec frame consists of log energy of the entire spectrum, log energy of the frequency region from 3 to 8 kHz, and a spectral flatness measure, all computed with a 5-msec window. A detected burst is considered to be correctly classified if it occurs within 20 msec of the closure-burst label boundary as obtained from the TIMIT phonetic alignments. Training was done on randomly-selected dialect regions of 40 sentences from the training partition of the TIMIT corpus, with 133 positive examples and 10760 negative examples. Testing was done on the test partition of one dialect region of the TIMIT corpus, using 320 sentences from 32 speakers. Niyogi et al. constructed ROC curves for their various classification methods, varying a parameter called U that "controls the trade-off between empirical fit to the data and capacity of the learning machine" [108]. In addition, they evaluated a phoneme-based HMM approach and a derivative-of-energy approach on the same data. Due to the large number of examples of non-bursts compared to the number of examples of bursts, an ROC curve in which performance is computed based on the total number of frames yields an extremely low number of false acceptances compared to the number of false rejections. To address this issue, Niyogi et al. constructed their ROC curves by evaluating the number of detected bursts with respect to the number of burst and non-burst *phonemes* instead of the number of burst and non-burst *frames*. Their ROC-curve results (with false-rejection rate on the X axis and correct-acceptance rate on the Y axis) are reproduced in Figure 6.10. It can be seen that the best performance is obtained by the non-linear SVM, with an equal-error rate of about 12% and a total error rate of 24%. The linear SVM has an equal-error rate of about 16%, the derivative-of-energy approach has an equal-error rate of about 20%, and the HMM approach has a total error rate of 36%.

6.5.2 Proposed Method for Impulse Detection

In our proposed method, we use knowledge of the physical processes involved in production of bursts in order to classify burst-related impulses. A burst is created by closure of the oral cavity in order to produce an increase in air pressure, which is followed by a sudden release of the constriction, causing an abrupt increase in energy of the signal. Because of

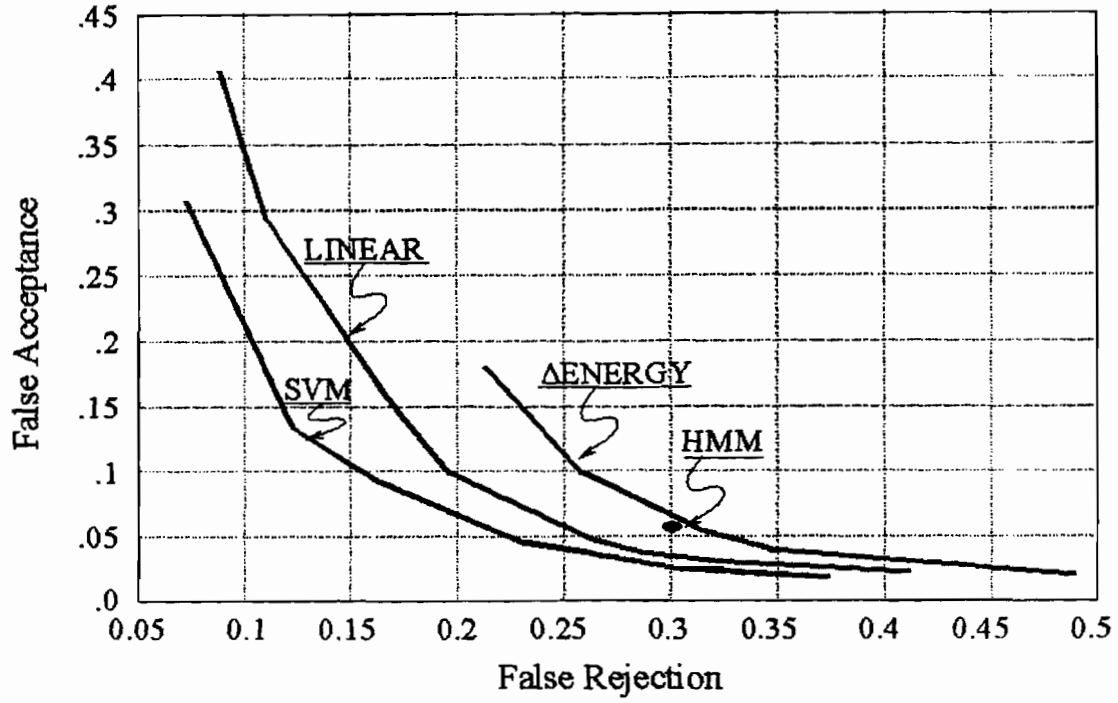

Figure 6.10: ROC curve of false rejection and false acceptance error rates for linear SVM (LINEAR), non-linear SVM (SVM), delta-energy (Δ ENERGY), and HMM (HMM) methods of impulse detection. Based on figure from Niyogi, Burges, and Ramesh, 1999 [108].

this process, bursts are characterized by about 15 to 30 msec of low energy (during the closure), which is followed by a sudden increase in energy (at the instant of release), which is followed by a gradual decline in energy (during the release). Furthermore, the radiation characteristic of sound emanating from the mouth causes the burst at the instant of release to take on the qualities of an impulse, with a relatively flat spectrum and short duration. Other research [133] has shown that the burst does not have a completely flat spectrum, but is shaped to some degree by the type of burst. The proposed method then detects burst-related impulses by applying the following criteria:

- There must be a relative increase in energy at the instant of release,
- The increase in energy must occur over most frequency bands, and

- The burst must have certain spectral properties that distinguish it from environmental noise (such as clicks).

These criteria can be satisfied by using the measure of intensity discrimination to estimate relative changes in energy, using the FLMP [100] to combine the frequency-band information into a single measurement, and using a neural-network classifier to incorporate spectral properties into the classification process.

The proposed method then works as follows:

1. Intensity discrimination is applied to bark-scale frequency bands. The window sizes for I and ΔI are small, in order to maximize the discrimination of impulses.

2. Equal-loudness weighting of the frequency bands is applied to the results of intensity discrimination, in order to give the frequency bands that are perceptually stronger greater weight in the final result.

3. Assuming independence of the frequency bands, the weighted results of intensity discrimination are combined using the FLMP. Each band is assumed to provide evidence for one of two conditions: a burst, or lack of a burst.

4. A threshold is used to select a number of "candidate bursts" for further processing.

5. A neural network is used to evaluate all candidate bursts, with input from the FLMP result and several frames of spectral information (PLP coefficients), and a binary burst/non-burst output.

Results from this process are illustrated in Figure 6.11, which shows an example waveform (containing letters of the alphabet) and its corresponding spectrogram, the higher six bands of intensity discrimination, the combined result of intensity discrimination, and the results of neural-network classification.

This method has two advantages over a derivative-of-energy approach. First, in most derivative-of-energy approaches (and in the SVM method), the absolute change in energy is computed; this absolute change can be influenced by "external" factors such as recording volume and speaking style. The use of intensity discrimination in the proposed method

normalizes for these sources of variability in a perceptually-motivated way. Second, a derivative-of-energy approach is unable to distinguish between burst-related impulses and impulses due to other factors. The proposed method accounts for the differences between these two types of impulses by taking into consideration the spectral properties of bursts and non-bursts. This method also provides an advantage over the SVM method in that the specific relationship between different frequency bands during an impulse is accounted for in the proposed method, while the SVM method relies on only two measures of energy at each frame. Finally, although support vector machines are thought to generalize well on test data, their run-time performance can be "abysmally slow" [17], which is a factor that must be taken into account if real-time processing is desired.

6.5.3 Implementation of Proposed Method

This method uses four parameters for locating candidate bursts: the window sizes for I and ΔI, the window size for computing the delta in ΔI, and the threshold value. Initial values for these parameters were determined from speech-specific knowledge and visual inspection of their effects. These initial values were then modified in small increments, and the resulting candidate bursts were evaluated on a development partition of the TIMIT corpus. Then, several sets of parameters with the best performance were selected for locating the candidate peaks on which to train neural networks. Finally, the set of parameters and the neural network with the best performance on the development-set data were selected as the final system.

Evaluation was done in the same way as Niyogi et al., with a detected impulse considered correctly detected if it lies within 20 msec of the manually-labeled closure-burst boundary, and the percentage of insertions and deletions measured relative to the number of plosive and non-plosive phonemes. The evaluation of the four parameters was done on the TIMIT corpus, training of the networks was done on the TIMIT, Stories, and Portland Cellular corpora, development-set evaluation of the networks was done on the TIMIT corpus, and test-set evaluation was done on the TIMIT, Stories, and Portland Cellular corpora. A variety of corpora were used in training the networks and in final evaluation, in order to build and evaluate a system on different channel conditions. Development-set

evaluation of parameters and networks was done on the TIMIT corpus, in order to have a single evaluation result on which to base the parameter and network selection.

The results of evaluation of the various parameter values are plotted in Figure 6.12, which shows the general trend for the set of all parameters. It can be seen that the number of insertions ranges from about 30% to about 75%, and the number of deletions ranges from 1.25% to 6%. As the neural network is trained only on the detected candidate bursts, insertion errors can be reduced by the network, but deletion errors can not be recovered from. This means that the final system will have more than 1.25% deletion errors, which we considered acceptable. The initial networks were trained with 13 PLP coefficients, no delta values, and a context window of frames at –5, 0, and 5 msec relative to the frame of interest. Once the final parameter set was chosen, then network training was changed to include delta values and a larger context window of frames from –30 to +30 msec relative to the frame of interest (at 5-msec intervals). Training was done on 2000 examples from the TIMIT corpus, 4000 examples from the Stories corpus, and 2000 examples from the Portland Cellular corpus; these values were selected to provide nearly equal quantities of positive and negative examples. The sets of parameter values used in training the neural networks and the resulting development-set performance of the networks are specified in Table 6.3. As in the baseline systems, training was done on 20,000 examples.

6.5.4 Results of Impulse Detection

From the development-set results, the parameter set of {I window size = 22.0, ΔI window size = 24.0, delta window size = 14.0, and threshold = 0.075} was selected for the final system. Test-set evaluation was done on 1344 sentences (6261 bursts and 45420 non-burst phonemes) from the TIMIT corpus, 42 sentences (2629 bursts and 19967 non-burst phonemes) from the Stories corpus, and 33 sentences (946 bursts and 8431 non-burst phonemes) from the Portland Cellular corpus. Results of these evaluations are given in Table 6.4.

It can be seen from this table that the total error rate on the TIMIT corpus is 13.20%, which is a 45% reduction in error compared to the best total error rate reported by Niyogi et al. on this corpus. Also, it is interesting to note that the increased noise,

Table 6.3: Sets of parameter values used for detecting candidate impulses, and performance of neural networks trained on the resulting candidate peaks (development-set results). The parameter set number in the left-most column can be used to locate the performance of this parameter set without neural-network classification by referring to Figure 6.12.

Parameter Set Number	I Window Size (msec)	ΔI Window Size (msec)	Delta Window Size (msec)	Threshold	Network Error (ins%+del%)
1	22.0	24.0	14.0	0.075	14.81
2	16.0	24.0	18.0	0.10	15.34
3	22.0	24.0	12.0	0.075	14.94
4	20.0	24.0	12.0	0.075	15.05
5	16.0	24.0	12.0	0.075	16.02
6	20.0	18.0	10.0	0.05	15.34
7	20.0	20.0	8.0	0.05	15.41
8	16.0	20.0	8.0	0.05	16.10

Table 6.4: Insertions, deletions, and total error rate for each of three corpora (test-set results).

Corpus	Insertions (%)	Deletions (%)	Total Error (%)
TIMIT	5.14	8.06	13.20
Stories	11.56	8.34	19.91
Portland Cellular	25.26	8.08	33.34

different channel conditions, and lower sampling rate did not greatly affect the deletion rate, but had a dramatic impact on the insertion rate. One possible explanation for this is that people change their speaking style to compensate for degraded channel conditions, thereby enunciating bursts clearly enough to be detected at roughly the same rate for any channel. The increased noise for the Stories and Portland Cellular corpora may explain the increased insertion rate, as these channels are more likely to have non-speech phenomena that resemble burst-related impulses. Finally, performance of this method (without optimizing the code in any way) is close to real-time on a Pentium Pro 200 MHz computer.

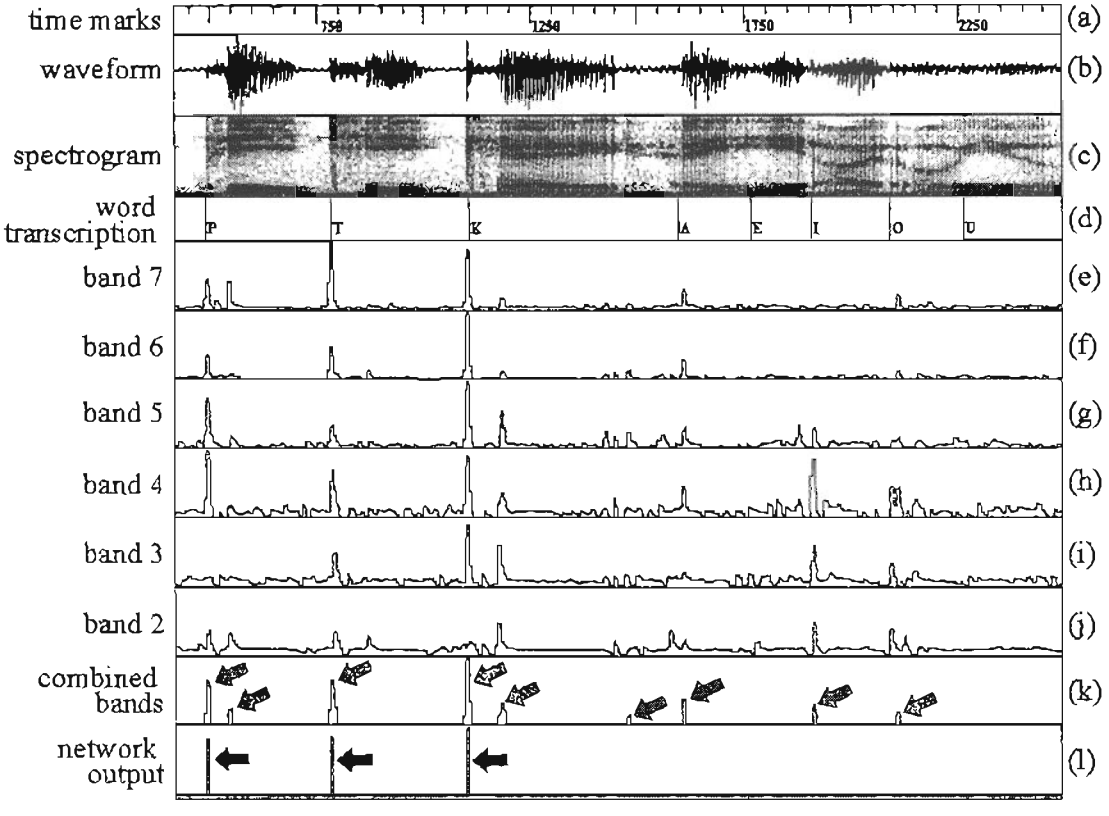

Figure 6.11: Illustration of proposed method of burst detection. The panels show (a) time marks, (b) the waveform for "P T K, A E I O U" uttered with background noise, (c) the spectrogram, (d) the word-level time-aligned transcription, (e)-(j) the intensity discrimination results for six of the frequency bands, (k) the result of combining the frequency bands using FLMP (with potential impulses indicated by arrows), and (l) the neural-network output of impulse detection (with detected impulses indicated by arrows).

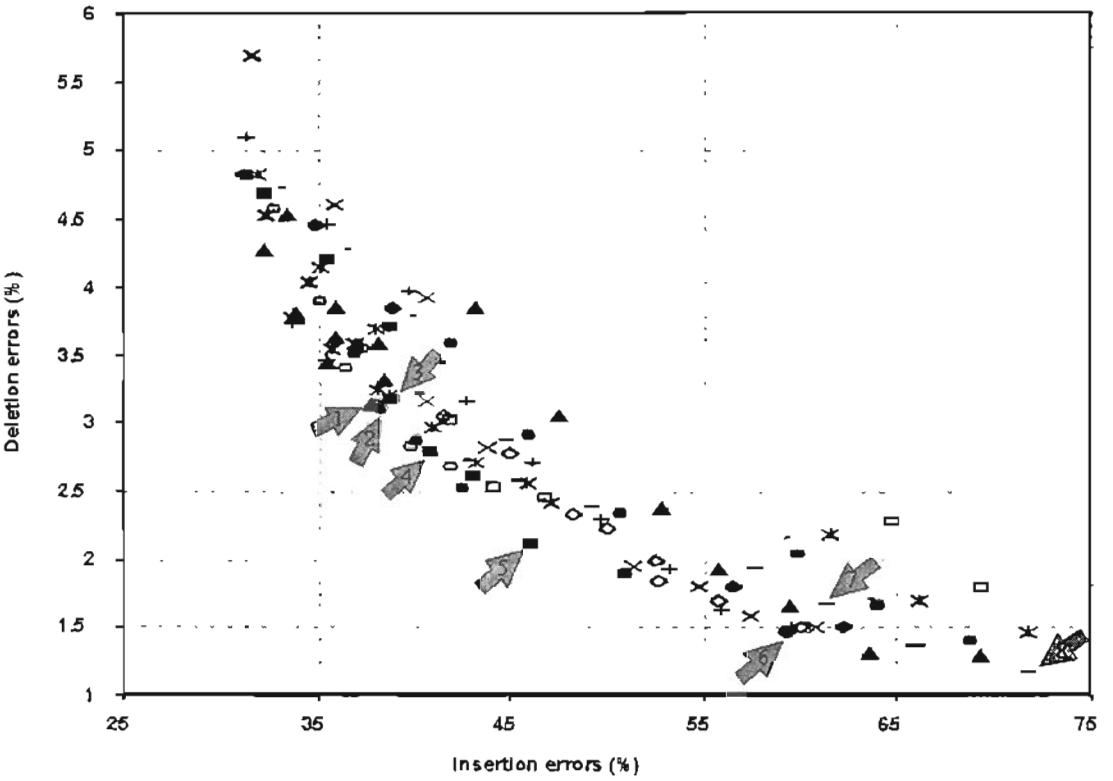

Figure 6.12: ROC curve of insertion errors and deletion errors for various parameters of the proposed impulse-detection method prior to neural-network classification. Points indicated by numbered arrows had their parameter values used in subsequent neural-network classification; these numbers can be used to determine the parameter values by referring to Table 6.3.

Chapter 7

Implementation of Distinctive Phonetic Features and Phonetic Transitions

As described in Chapter 5, the system proposed in this thesis combines phonetic transition information and distinctive phonetic features to better utilize acoustic-phonetic information. Moreover, we have argued that some of this information is not modeled effectively by current HMM systems. In this section, we describe how the distinctive phonetic features and phonetic transitions have been implemented in our proposed alignment system.

7.1 Set of Distinctive Phonetic Features

The distinctive phonetic features used for this thesis are Manner, Place, and Height. This set was chosen to be as small as possible in order to better assume independence of the values (a larger set may have necessitated redundant features), while being large enough to uniquely specify 43 English phonemes as well as pauses and breath noise. In some linguistic theories, distinctive phonetic features are restricted to have binary values, such as [+voice] and [–voice]. As Ladefoged notes, however, when the values of a feature are mutually exclusive (such as [labial], [alveolar], and [dorsal] for the Place feature), it can be more natural to allow the features to take on multiple values. Furthermore, the use of a set of mutually exclusive values in a single neural network allows discriminative training, which is not easily possible if a set of binary-valued networks are created. Because of these advantages, and the desire to use a small number of independent features, we use multi-valued features.

The features, their values, and the linguistic meaning associated with these values are given in Table 7.1. These values are based on Ladefoged [85], and have been modified in order to achieve a set of 35 unique phonemes (not including diphthongs, silences, or breath noise), as specified in Table 7.2. As an example of a modification for this thesis, Ladefoged states that all consonants except for /w/ and /j/ are of maximum height, and yet based on preliminary results from neural-network outputs we found that the consonants /l/ and /ɻ/ are better represented by height values of [h2] and [h4], respectively. In addition, the diphthongs and affricates are represented by their component phonemes, as specified in Table 7.3. The phonemes /h/, /ɦ/, and /.br/ are affected strongly by their context, and so the assigned Place value of [unk] ("unknown") is given a value at run-time of the highest probability from the set of Place values [fnt], [mid], and [bck]. This assignment is made based on the assumption that the current phoneme has a place of articulation similar to its surrounding vowels. A Height value of [unk] (for /.br/) is assigned a default value of [max]. Although 43 phonemes (including the eight English diphthongs and affricates) can be identified, discrimination is not possible for the following phonemes (listed in Table 7.4): /ɔ/ and /ɑ/, the retroflex sounds /ɻ/ and /ɝ/, and the flaps /ɾd/, /ɾt/, and /ɾn/. For the purposes of phonetic alignment, distinguishing between these phonemes is only important when they occur in sequence, such as in the word "jurors" (/ʤ ʊ ɻ ɝ z/). In such cases, we rely on the phonetic transition classifier to determine at which time point the transition between the two phonemes occurs.

While the proposed set of distinctive phonetic features was designed for American English, we hope that only minor modifications will be needed to extend this set to be useful for many of the world's languages. For example, the addition of a Rounding feature may be necessary for German, or a [trill] value may be added to the Manner feature for distinguishing the Spanish /r/ phoneme from the American English /ɻ/. In theory, however, even some phonemes that do not occur in American English can be specified in terms of the existing set of features. As an example, the voiced alveolar approximant /ɹ/ (in some British-English variants of the word "red" /ɹ ɛ d/ [86]) can be represented with a Manner value of [approximant], a Place value of [alveolar], and a Height value of [h2].

Table 7.1: Values for each of the three distinctive phonetic features used in this study.

Features	**Value**	**Meaning**
Manner	vow	vowel
	app	approximant
	nas	nasal
	asp	aspiration
	frc	fricative
	vfr	voiced fricative
	stp	stop (plosive)
	vst	voiced stop (voiced plosive)
	flp	flap
	bre	breath noise
	clo	closure
Place	fnt	front
	mid	mid
	bck	back
	ret	retroflex
	lat	lateral
	lab	labial
	den	dental
	alv	alveolar
	dor	dorsal
	clo	closure
Height	max	maximum height
	h1	very low height
	h2	low height
	h3	high height
	h4	very high height
	clo	closure

Table 7.2: Phonetic Symbols in Worldbet (column 1) and IPA (column 2), and their corresponding Manner, Place, and Height values (columns 3, 4, and 5). Examples with each phoneme in an English word are given in the final column.

Worldbet Symbol	IPA Symbol	Manner	Place	Height	Example Word
i:	iː	vow	fnt	h4	b<u>ee</u>t
i	i	vow	fnt	h4	(*used in diphthong*)
I	ɪ	vow	fnt	h3	b<u>i</u>t
E	ɛ	vow	fnt	h2	b<u>e</u>t
@	æ	vow	fnt	h1	b<u>a</u>t
u_x	ʉ	vow	mid	h4	s<u>ui</u>t
I_x	ɨ	vow	mid	h3	rose<u>s</u>
^	ʌ	vow	mid	h2	ab<u>o</u>ve
&	ə	vow	mid	h2	<u>a</u>bove
u	u	vow	bck	h4	b<u>oo</u>t
U	ʊ	vow	bck	h3	b<u>oo</u>k
o	o	vow	bck	h2	(*used in diphthong*)
>	ɔ	vow	bck	h1	c<u>augh</u>t
A	ɑ	vow	bck	h1	f<u>a</u>ther
a	a	vow	bck	h1	(*used in diphthong*)
j	j	app	fnt	h4	<u>y</u>es
w	w	app	bck	h4	<u>w</u>ent
l	l	app	lat	h4	<u>l</u>ent
l=	l̩	app	lat	h4	bottl<u>e</u>
9r	ɻ	app	ret	h2	<u>r</u>ent
3r	ɝ	app	ret	h2	b<u>ir</u>d
&r	ɚ	app	ret	h2	butt<u>er</u>
&_0	ə̥	asp	mid	h2	t<u>o</u> go
h	h	asp	unk	max	<u>h</u>ope
h_v	ɦ	asp	unk	max	she <u>h</u>ad
m	m	nas	lab	max	<u>m</u>e
n	n	nas	alv	max	<u>kn</u>ee
N	ŋ	nas	dor	max	sing
m=	m̩	nas	lab	max	bottom
n=	n̩	nas	alv	max	button
N=	ŋ̩	nas	dor	max	increasing your
ph	p^h	stp	lab	max	<u>p</u>an
th	t^h	stp	alv	max	<u>t</u>an
kh	k^h	stp	dor	max	<u>c</u>an

continued from previous page

Worldbet Symbol	IPA Symbol	Manner	Place	Height	Example Word
b	b	vst	lab	max	ban
d	d	vst	alv	max	dan
g	ɡ	vst	dor	max	gander
d_(	ɾd	flp	alv	max	rider
th_(	ɾt	flp	alv	max	writer
n_(	ɾn	flp	alv	max	winner
f	f	frc	lab	max	fine
T	θ	frc	den	max	thigh
s	s	frc	alv	max	sign
S	ʃ	frc	fnt	max	shine
v	v	vfr	lab	max	vine
D	ð	vfr	den	max	this
z	z	vfr	alv	max	resign
Z	ʒ	vfr	fnt	max	azure
pc	—	clo	clo	clo	_pan
tc	—	clo	clo	clo	_tan
kc	—	clo	clo	clo	_can
bc	—	clo	clo	clo	_ban
dc	—	clo	clo	clo	_dan
gc	—	clo	clo	clo	_gander
tSc	—	clo	clo	clo	_church
dZc	—	clo	clo	clo	_judge
.pau	—	clo	clo	clo	—
.br	—	bre	unk	unk	—

Table 7.3: Diphthongs and their corresponding component phonemes, as used in the proposed system.

Diphthong/Affricate	Left Phoneme	Right Phoneme
ei	E	i
aU	a	U
oU	o	U
aI	a	I
>i	>	i
iU	i	u
tS	th	S
dZ	d	Z

Table 7.4: Phonemes not distinguished by proposed distinctive-feature set, and their minimal pairs

Phoneme	Minimal-Pair Word	Word Pronunciation
ɔ	caught	k ɔ t
ɑ	cot	k ɑ t
ɻ	free	f ɻ i:
ɝ	furry	f ɝ i:
ɾd	wider	w aɪ ɾd ɚ
ɾt	whiter	w aɪ ɾt ɚ
ɾn	whiner	w aɪ ɾn ɚ

7.2 Combining Distinctive Phonetic Features

Given the set of distinctive phonetic features described in the previous section, the remaining issue is how the values of these features will be combined to produce a phoneme-level representation. In the proposed system, we construct context-dependent steady-state phonemes from the distinctive feature information using a combination of both context-dependent and context-independent distinctive features. The context-dependent framework is the same as in our baseline phonetic recognition system, in that a given phonetic category can be dependent on the context of the preceding phoneme, be dependent on the context of the following phoneme, or be context independent.

For the Manner feature, context-independent categories are used, because in general the manner of articulation is not greatly influenced by the preceding or following phonemes (with the exceptions of approximants and vowels in the context of approximants or vowels). The resulting network has 11 categories, one for each type of manner of articulation.

For the Place feature, both context-dependent categories and context-independent categories are used. The closure, labial, dental, alveolar, and dorsal categories are considered context independent, and other categories (front vowel, mid vowel, back vowel, retroflex, and lateral) are context dependent. Currently, the distinction between context-independent and context-dependent categories is based on the presence or absence of

formant values, although in future alignment systems all Place categories may be context-dependent. The context values are also based on place of articulation, and so there are 10 possible contexts. In determining the context of a plosive closure, the closure is mapped to the place of articulation of its corresponding plosive. So, for example, a front vowel in the context of a following /p^h/-closure is represented as "fnt$\succ$lab" (where $\succ$ indicates transition), and a front vowel in the context of a following /t^h/-closure is represented as "fnt$\succ$alv." This allows classification of unreleased plosives based on the closure alone, using information in the surrounding phonemes' resonant-frequency trajectories. The resulting network has 108 output categories.

For the Height feature, context-independent categories are used in order to simplify the construction of context-dependent phonetic-level categories (with context based solely on place of articulation), although future alignment systems may use context-dependent Height categories. In order to account for the fact that the property of Height may be influenced by coarticulation (thus implying an advantage to context-dependent categories), a large number of hidden nodes are used in this network, as in the context-independent phonetic-level "big dumb neural network" systems proposed by Bourlard and Morgan [10]. The resulting network has 300 hidden nodes and six output categories (one for each Height value).

The Manner, Place, and Height categories are combined to arrive at a phoneme-level representation that is context-dependent (as illustrated in Figure 5.2), where the context is based on the place of articulation of the neighboring phoneme. The values of each feature are combined using the Fuzzy Logic Model of Perception (FLMP), which assumes independence between each distinctive feature. As suggested by Zadeh and reported in Oden and Massaro [111], we use exponential weighting factors to adjust the relative importance of each feature in its contribution to the resulting phoneme. The values for these weighting factors have been determined empirically by an iterative process of modifying a weight and evaluating its effect on alignment accuracy, until further improvement on the development set is not obtained.

7.3 Implementation of Phonetic Transition Information

The phonetic transition categories are derived using the distinctive features described in Section 7.1. As in the steady-state networks, the distinctive-feature transition networks have their outputs combined to arrive at phoneme-level transition probabilities. The outputs of each transition network are the possible combinations of values for each distinctive feature, as well as a single "non-transition" category. For the Manner transition network (with 11 values for Manner) there are 122 outputs, for the Place transition network (with 10 values for Place) there are 101 outputs, and for the Height transition network (with 6 values for Height) there are 37 outputs.

As an example of how these networks are combined to determine phoneme-level transition probabilities, we can consider the transition from /p^{h}/ to /ɑ/, as in the word "pot." The /p^{h}/ phoneme has the values [stp] for Manner, [lab] for Place, and [max] for Height, and /ɑ/ has the values [vow] for Manner, [bck] for Place, and [h1] for Height. The probability of a transition from /p^{h}/ to /ɑ/ given a certain observation is then the combination of the probabilities of "stp$\succ$vow", "lab$\succ$bck", and "max$\succ$h1" for that observation. Considering these probabilities to be equivalent to "fuzzy truth values", we can then use the FLMP to arrive at a final probability of transition from /p^{h}/ to /ɑ/:

$$p(/\mathrm{p}^{\mathrm{h}}/ \succ /\mathrm{ɑ}/|o) = \frac{p(\mathrm{stp} \succ \mathrm{vow}|o) \cdot p(\mathrm{lab} \succ \mathrm{bck}|o) \cdot p(\mathrm{max} \succ \mathrm{h1}|o)}{\sum_X \sum_Y p(\mathcal{M}(X) \succ \mathcal{M}(Y)|o) \cdot p(\mathcal{P}(X) \succ \mathcal{P}(Y)|o) \cdot p(\mathcal{H}(X) \succ \mathcal{H}(Y)|o)} \tag{7.1}$$

where $\succ$ indicates transition, X and Y are phonemes, $\mathcal{M}(X)$ is the Manner of phoneme X, $\mathcal{P}(X)$ is the Place of phoneme X, and $\mathcal{H}(X)$ is the Height of phoneme X. As in the steady-state estimation, empirically-derived exponential weights are used to adjust the relative importance of each feature, although these weights are not indicated in Equation 7.1.

Given the within-phoneme ("steady-state") probabilities and the phonetic transition probabilities for each observation, we can then estimate the most likely phonetic sequence by combining these probabilities during the Viterbi search. Combination of the probabilities during the search allows relatively early integration of the within-phoneme and transition information. (A comparison with other systems that combine transition and within-phoneme probabilities is given in Section 7.5.)

7.4 Training Issues

The data from manually time-aligned phonetic transcriptions were used to select the within-phoneme and phonetic transition training samples. First, the same mapping procedure that was used in the baseline system was applied to remove diacritics and short pauses. Then, phonetic labels were mapped to their respective distinctive features. For training the transition networks, the region of ±5 msec from each phonetic boundary was marked as a transition region. With a frame rate of 5 msec, this yielded two training samples for every phonetic boundary.

The corpora used for training were the TIMIT [71], Stories [27], and Portland Cellular [27] corpora (the same corpora and data files that were used in the baseline recognition system). They were selected to provide a variety of channel conditions for training. The speech data were converted to a 16 kHz sampling rate, if necessary, and high-pass filtered at 160 Hz before computing the PLP features, in order to make the microphone-speech training data more closely match the telephone-speech data. Each network was trained using the same 5-msec frame size as in the baseline system.

For the within-phoneme classification networks, the feature set consisted of PLP features with the same 120-msec context window used in the baseline system, as well as the acoustic-level features thought to be relevant to the particular distinctive feature. For the transition networks, the same PLP features were used, but with a more narrow 60-msec context window. Table 7.5 lists the acoustic-level features used as input to each type of network.

For the within-phoneme ("steady-state") networks, F0 values were provided to the network to give information about the speaker's gender. Because F0 is highly correlated with gender, and because gender correlates with vocal tract length and thus influences the locations of formant frequencies, F0 values may allow the network to better learn the formant frequencies associated with Manner, Place, and Height. Voicing and voice-onset time provided the Manner network with information about voicing, and provided the Place and Height networks with information about the location of the current frame with respect to any surrounding consonants. Such consonants may have a coarticulatory

Table 7.5: Acoustic-level features used in each type of distinctive-feature network.

Network	Acoustic-Level Features
Steady-State Manner	F0, voicing, voice-onset time, glottalization, burst-related impulses, and intensity discrimination
Steady-State Place	F0, voicing, voice-onset time, and intensity discrimination
Steady-State Height	F0, voicing, voice-onset time, and intensity discrimination
Phonetic Transition	F0, delta-F0, voicing, voice-onset time, glottalization, burst-related impulses, and intensity discrimination

effect on the current frame; the closer the current frame is to the consonant, the stronger the coarticulatory effects can be.

These PLP and acoustic-level features were input to a fully connected feed-forward neural network, which was trained using back-propagation to estimate the likelihood of each category. The training was adjusted to use the negative penalty modification proposed by Wei and van Vuuren [145], as in the baseline system, and training was stopped after 45 iterations. During the Viterbi search, the baseline method of applying duration limits using penalties was used. The network results for iterations 15 through 45 were applied to the forced-alignment task on a development set of the TIMIT corpus, and the "best" iteration was determined by selecting the iteration with the minimum alignment error.

7.5 Comparison with Previous Work

If we compare the number of categories required to compute context-dependent phonetic probabilities using distinctive phonetic features (125) with the number of categories required for a context-dependent phoneme-level classifier (about 600), the distinctive-feature approach uses about one-fifth the number of categories. If we compare the number of categories required to compute transition information using distinctive phonetic features and the FLMP (260) with the number of categories required to compute transition information using only phoneme-level information (1444 for a set of 38 phonemes), the distinctive-feature approach again requires about one-fifth the number of categories. This reduction

is possible because of the explicit model for combining the distinctive-feature information. The result of this reduction in the number of categories is that for a fixed amount of training data, there is a much greater number of training samples per category with the distinctive-feature approach than with the baseline phonetic approach. This is especially advantageous when using neural networks for classification, as learning tends to be poor for categories that have a relatively small number of training samples.

In the proposed system, transition information that depends on the observed speech frame is incorporated with the within-phoneme probabilities during the Viterbi search; this allows both acoustics-based transition probabilities and context-dependent phonetic probabilities to influence the most likely path through the HMM. In contrast, the SUMMIT system utilizes the transition information after the search has determined the most likely sequence of segments. The SUMMIT approach may be helpful in estimating overall word likelihoods, but the transition information does not influence the segmentation boundaries. The SPAM system uses the Viterbi search for both within-phoneme and phonetic transition categories, but combines the within-phoneme and transition results at the word level, after each Viterbi search. This late-integration approach is unable to take into account the timing relationship between phonemes and phonetic transitions. In the diphone-based system developed at OGI prior to this thesis [67], the constraint that phonetic steady-state and transition categories must occur in alternating order is enforced during the Viterbi search, but the state occupation probability does not depend on transition probabilities.

The proposed system uses context-dependent within-phoneme categories, which is made possible by the use of separate within-phoneme and phonetic transition networks. The OGI diphone [67] and CSELT [34] systems use context-independent categories, in order to improve the number of training samples per category in the single classifier. In addition, in order to keep the number of categories as small as possible, the OGI diphone system was trained only on the small-vocabulary digits task, and the CSELT system has a reduced set of transition and phonetic categories.

In the Discriminant HMM approach [10], the previous state is included in the estimate of the current state. This requires the classifier to have state information as an

input, and Bourlard reported that the resulting complexity prevented the full Discriminant HMM approach from being implemented in practice [10]. The IOHMM approach [8] requires a third "emit-or-not" distribution in addition to the state occupation and transition probabilities; this extra distribution is not required in the proposed system. In the HNN approach [127], two networks are trained for each *category*; one to estimate the state occupation probability, and the other to estimate the state transition probability. It may be because of this large number of required networks that the final HNN system uses the traditional HMM approach to state transitions instead of observation-dependent transition probabilities.

The proposed system is the only known system in which the FLMP is used to combine distinctive features to arrive at a phoneme-level probability estimation. Other systems that employ distinctive features use the distinctive-feature results as input to a phoneme-level GMM or ANN classifier [76, 82, 69, 131, 36], or require a large number of combinations of distinctive-feature HMM states [41, 44].

Chapter 8

Evaluation Methodology

8.1 Agreement

For evaluating the relative performance of the baseline and proposed automatic alignment systems, we measure both agreement with manual alignments and robustness. In addition, we investigate the performance of several aspects of the proposed system, and conduct tests to evaluate its usefulness in practical applications.

Agreement with manual alignments is measured for a set of 13 corpora. Judgements about the quality of alignment agreement are determined based on the commonly-used threshold of 20 msec, but we also present results for other thresholds. Significance of the difference between the baseline and proposed methods is computed for each corpus using McNemar's test, with a significance level of 0.05. In addition, we compare the results of the baseline and proposed methods with levels of inter-labeler agreement.

8.2 Robustness

In measuring robustness, we evaluate both systems on the many variations of the TIMIT corpus. The data for the original TIMIT corpus were recorded with a close-talking noise-canceling head-mounted Sennheiser microphone (model HMD-414) [53] and digitized at 16 kHz, and so the speech data in this corpus are of high acoustic quality. Other variations of this corpus have been created; each (except for the FFM-TIMIT corpus) was created by playing the original TIMIT speech through a speaker, recording it via some telephone or microphone channel, and shifting the resulting speech in the time domain so that

the channel-distorted speech is aligned with the original TIMIT speech. In this way, the phonetic labels that correspond to the original TIMIT speech also correspond to the TIMIT variants. The FFM-TIMIT corpus was recorded simultaneously with the recording of TIMIT, using a Breul & Kjaer 1/2" free-field microphone (model 4165). The label alignments of FFM-TIMIT were then increased by 1.25 msec to account for the distance (time delay) between the Sennheiser and Bruel & Kjaer microphones. As a result, the FFM-TIMIT corpus has microphone characteristics different from TIMIT, but the data have not been subjected to post-processing loudspeaker and channel distortion.

Because the phonetic alignments of the speech in the TIMIT variants are the same but the recording conditions differ, differences in automatic-alignment results, when evaluated on these corpora, must be due to the effect of the channel and loudspeaker conditions. For any given boundary, we can measure the standard deviation of alignments for that boundary, using either the proposed or baseline alignment systems. For example, if we evaluate the baseline system on a boundary between /.pau/ and /ʃ/ in a given file, there will be one result from each TIMIT corpus, for a total of 12 results. The standard deviation of these 12 results can be computed, providing a measure of the robustness of the baseline system on that phonetic boundary. We can then compute an average standard deviation for both systems, and evaluate each system's robustness with respect to a change in channel conditions.

An advantage of this approach to measuring robustness is that it does not rely at all on the manual boundary information. A disadvantage of this approach is that the change in channel conditions can increase the signal-to-noise ratio and obscure certain phonetic events. If such channel conditions were present during a real communication, we may assume that the speaker would change his or her speaking style to improve the communication of important events in the speech signal (known as the Lombard effect). With the various TIMIT corpora, the speaking style is not altered, and it is possible that the increased noise makes the detection of certain acoustic-phonetic events impossible. In cases with a low signal-to-noise ratio, we would predict that our proposed method has no advantage over the baseline system, because the proposed method relies on extracting acoustic-phonetic cues for improved performance. (We would also expect, under such

conditions, a noticeable decrease in the consistency of manual alignments. Such a decrease was found in the manual alignments of NTIMIT and CTIMIT reported in Section 4.2.) If the speaking style were altered in noisy environments to emphasize the phonetically-relevant aspects (which would occur in normal communication), we expect our proposed system to have better performance than the baseline system. We then note that under "real-life" circumstances with these various channels, the proposed method may be more accurate than is reflected in the observed results from TIMIT variants.

8.3 Other Issues

Once we have evaluated the relative performance of the two systems, it is important to identify which acoustic-phonetic features contribute to performance gains, and so we evaluate the alignment of specific phonetic boundaries for which we expect acoustic-phonetic features to improve performance. Specifically, we evaluate the alignment of voiced-unvoiced boundaries, boundaries between voiced consonants and vowels (for which F0 information may be useful [129]), boundaries between glottalized phonemes, silence-fricative boundaries (for which intensity discrimination may be useful), and closure-burst boundaries. These tests evaluate the effectiveness of the entire system, including phonetic transitions, distinctive features, and acoustic-level features, on specific types of boundaries. In addition, we evaluate the proposed system without the use of phonetic transition probabilities to determine the effect of transition information separately from distinctive features or acoustic-level features. Finally, we train a baseline system with acoustic-level features as input to the neural network, to evaluate the effectiveness of these features without the use of transition information or distinctive features.

Having measured the agreement with manual alignments and robustness of the baseline and proposed systems, questions remain about the usefulness of the proposed method. First, is the proposed method notably faster than manual alignment? If not, then the benefits of this method currently apply only to cases in which the use of human labelers is impractical or prohibitively expensive. To answer this question, we measure the execution time of the proposed method, and compare this with the time required for manual

alignment. Second, are the results of the proposed alignment system sufficiently different from the baseline system to have an effect on applications that use automatic phonetic alignment? A positive answer to this question indicates that the proposed system is of practical importance. For this reason, we train and evaluate a recognizer on the task of alpha-digit recognition (recognition of the letters of the alphabet, the ten digits from zero through nine, and the digit "oh"). We use the OGI Alphadigits corpus, which contains telephone-band speech of continuously-spoken letters and digits, with six letters or digits per utterance. As this corpus has no manual phonetic labels, automatic alignment is necessary in order to train an HMM/ANN alpha-digit recognizer on this corpus. We label the training data with both the baseline and proposed alignment methods (using canonical pronunciations for each letter and digit), and train separate recognizers on each set of labels. Test-set evaluation is performed, and the two systems are checked for statistically significant differences using McNemar's test, with a significance level of 0.05.

8.4 Summary

In summary, we evaluate the baseline system and proposed system using a measure of consistency with manual alignments and a measure of robustness. We base the final decision of success of this method on whether or not the proposed system shows better performance than the baseline system in both of these measures, although individual researchers are encouraged to form their own conclusions from the data. We investigate the effectiveness of various aspects of the proposed method. We also report on the execution time required by the proposed method, as it should be notably faster than manual alignment in order to be of practical use. We further evaluate the usefulness of the proposed method by training and evaluating two alpha-digit recognition systems, trained on labels generated from the baseline and proposed methods, respectively.

Chapter 9

Results and Discussion

9.1 Agreement with Manual Alignments

The 14 corpora used in evaluating agreement levels are described in Table 9.1. The results for the baseline and proposed alignment systems on these corpora are presented in Table 9.2. The relative reduction in error between the baseline and proposed systems is shown in Figure 9.1, according to the formula

$$\frac{E(baseline) - E(proposed)}{E(baseline)} \times 100\% \tag{9.1}$$

where $E(baseline)$ is the percent error (disagreement) of the baseline system with a 20-msec threshold, and $E(proposed)$ is the error (disagreement) of the proposed system with a 20-msec threshold. The average reduction in error across all 14 corpora is 27.92%, with a minimum reduction of 20.36% for Switchboard and a maximum reduction of 41.42% for the MWM Diphone corpus. Based on a 20-msec threshold, all of the results from the proposed system are significantly better than the baseline results, using McNemar's test with a significance level of 0.05 (all p values are less than 10^{-5}).

One reason why the reduction in error on Switchboard is the lowest of the corpora that we tested may be because the phonetic labeling of this corpus does not distinguish between the closure and burst portions of plosive phonemes. Because there is no labeling of the instant of burst onset, we combine the closure and burst outputs from the forced-alignment systems when evaluating on this corpus. As the proposed system has a large reduction in error over the baseline system at labeling closure-burst boundaries (see Table 9.7) the lack of this type of boundary reduces the relative effectiveness of the proposed system.

Table 9.1: Corpora used in comparing performance of the baseline and proposed methods to manual alignments.

Corpus	Description
TIMIT	Speech recorded with a head-mounted, noise-canceling microphone in a clean environment. The corpus consists of read, phonetically-balanced sentences. Over 1,300 files are available for testing.
Stories	Telephone speech from 688 people across the United States. Each file contains extemporaneous speech of about 1 minute in length. Over 200 files have been phonetically transcribed and time-aligned, and 40 of these files are used for test-set evaluation.
Portland Cellular	Cellular-telephone speech from 515 people in the Portland, Oregon area. Each file contains extemporaneous speech of about 1 minute in length. 200 files have been phonetically transcribed and time-aligned, and 33 of these files are used for test-set evaluation.
FFM-TIMIT	Recorded at the same time as TIMIT, using a Breul & Kjaer free-field microphone. Significant amounts of very-low-frequency noise.
Switchboard	Cellular-telephone speech from a large number of people on various topics. Phonetic labeling of this corpus has been provided by Joe Picone. Over 1,200 files are available for testing, although the files are of different lengths.
Kids' Speech	Head-mounted microphone speech recorded onto computer with a SoundBlaster^{TM} audio card. The speech was collected from children in grades K through 10. Isolated, prompted words from grades 1 through 10 have been phonetically aligned, and 168 of these files are used in this evaluation.
Names	Telephone speech of 30,000 first and last names from people across the United States. Over 6,500 utterances have been transcribed and time-aligned at the phoneme level, and 1,280 files are used in this evaluation.
Numbers	Telephone speech of various numbers, such as ZIP codes or address numbers. Collected from a large number of people across the United States. Over 6,500 utterances have been transcribed and time-aligned at the phoneme level, and almost 1,300 files are used in this evaluation.
Spelled and Spoken Words	Telephone speech of locations, people names, the alphabet, and spelled names. Collected from 4,000 people across the United States.

continued from previous page

Corpus	Description
Multi-Language Telephone Speech (MLTS), German	Telephone speech from 189 people speaking extemporaneous German speech for about 1 minute. Of these files, 104 have been manually transcribed and aligned at the phoneme level. Of these 104 files, half of the ones for which all phonemes can be identified (no ".nitl" or non-standard phonemes) are used in this evaluation. This results in test-set 30 files. Non-English phonemes were mapped to their closest English representation.
Multi-Language Telephone Speech (MLTS), Spanish	Telephone speech from almost 200 people speaking extemporaneous Spanish speech for about 1 minute. Of these files, 108 have been manually transcribed and aligned at the phoneme level. Of these 108 files, half of the ones for which all phonemes can be identified (no ".nitl" or non-standard phonemes) are used. This results in 30 test-set files for evaluation. Non-English phonemes were mapped to their closest English representation.
Multi-Language Telephone Speech (MLTS), Mandarin	Telephone speech from over 160 people speaking extemporaneous Mandarin speech for about 1 minute. Of these files, 70 have been manually transcribed and aligned at the phoneme level. Of these 70 files, half of the ones for which all phonemes can be identified (no ".nitl" or non-standard phonemes) are used. This results in 15 test-set files for evaluation. Non-English phonemes were mapped to their closest English representation.
Multi-Language Telephone Speech (MLTS), Japanese	Telephone speech from 161 people speaking extemporaneous Japanese speech for about 1 minute. Of these files, 64 have been manually transcribed and aligned at the phoneme level. Of these 64 files, half of the ones for which all phonemes can be identified (no ".nitl" or non-standard phonemes) are used. This results in 24 test-set files for evaluation. Non-English phonemes were mapped to their closest English representation.
MWM Diphones	High-quality microphone speech of diphones in nonsense carrier phrases, from a single American speaker. In evaluating alignment methods on this corpus, only the boundary of the target diphone was used when computing the agreement score.

Table 9.2: Percent agreement with manual alignments for the baseline and proposed methods, and relative reduction in error for the proposed method. The TIMIT, Stories, and Portland Cellular corpora were the corpora used in training and development.

Corpus	**Baseline Agreement (%)**	**Proposed Method Agreement (%)**	**Relative Reduction in Error (%)**
TIMIT	89.95	92.57	26.07
FFM-TIMIT	87.91	90.62	22.42
Stories	87.35	90.24	22.85
Portland Cellular	82.51	88.89	36.48
Switchboard	81.24	85.06	20.36
Kids	60.38	74.02	34.43
Names	82.00	86.67	25.94
Numbers	82.19	85.96	21.17
Spelled & Spoken	71.03	80.11	31.34
MLTS-German	78.84	87.49	40.88
MLTS-Mandarin	73.41	79.33	22.26
MLTS-Spanish	77.71	82.82	22.93
MLTS-Japanese	78.65	83.41	22.30
MWM Diphone	72.48	83.88	41.42

The Numbers corpus may have lower-than average reduction in error because of the small number of bursts in the vocabulary, and because the boundary between /oʊ/ and /ɹ/ in the word "four" is often indistinct, making manual alignment difficult. The MWM Diphone corpus may have relatively better performance because of the clear articulation of diphones that accentuates the acoustic-phonetic properties of the speech.

The agreement levels for the baseline and proposed system on all corpora at thresholds from 10 msec through 50 msec (at 10-msec intervals) are given in Tables 9.3 and 9.4, respectively. The results of the baseline and proposed system are compared with inter-labeler agreement in Table 9.5. It can be seen that manual labeling yields, on average, a 41.9% reduction in error over the baseline system and a 13.5% reduction in error over the proposed system. It is also interesting to note that for two of the five corpora in this evaluation, the proposed method has somewhat better agreement with the reference labels than a human labeler.

The closeness of the proposed method's results to the levels of inter-labeler agreement is encouraging, although the results for Mandarin speech indicate that there are corpora for which the proposed method does considerably worse than manual alignment. The poor results for Mandarin may result from the coarse mapping of Mandarin to English phonemes, in which it is difficult to assign a single English phoneme to a target Mandarin phoneme.

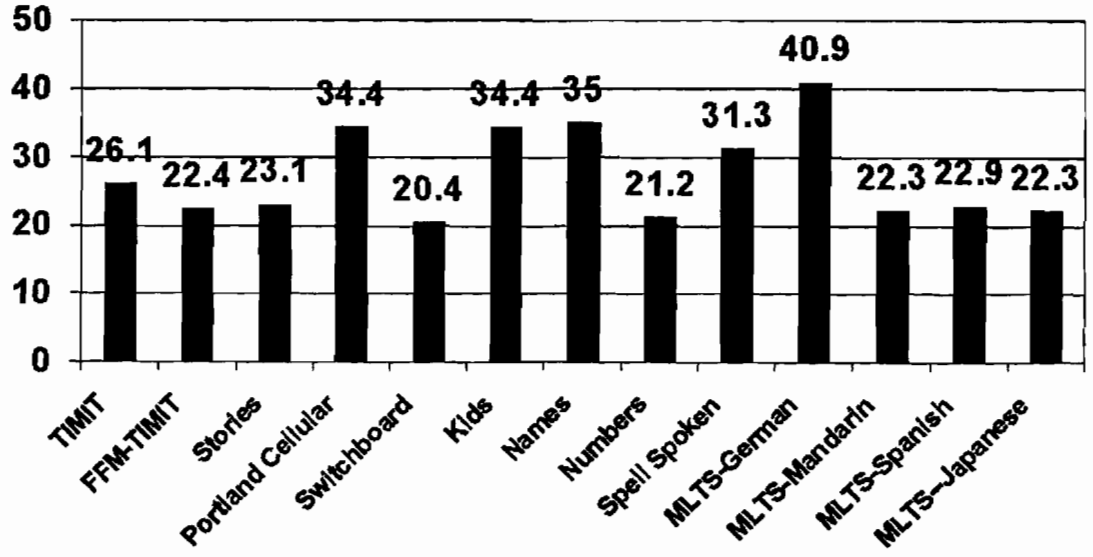

Figure 9.1: Graph of relative reduction in error from the baseline to the proposed method, for each corpus.

9.2 Robustness Measurements

The robustness of the baseline and proposed systems was evaluated by computing the standard deviation of results for each phonetic boundary as evaluated on 12 TIMIT corpora, and then finding the average standard deviation of all boundaries. The TIMIT corpora used in this evaluation are described in Table 9.6. Because the waveforms of the speech files in the original HTIMIT corpus may be shifted by up to 50 msec [125], we used a version of HTIMIT developed by Sarel van Vuuren that has the speech data more

Table 9.3: Agreement levels for the baseline system on all corpora at several thresholds.

Corpus	Agreement within 10 msec (%)	Agreement within 20 msec (%)	Agreement within 30 msec (%)	Agreement within 40 msec (%)	Agreement within 50 msec (%)
TIMIT	72.95	89.95	95.21	97.39	98.45
FFM-TIMIT	69.03	87.91	93.92	96.52	97.85
Stories	70.77	87.35	93.27	95.98	97.27
Portland Cellular	65.04	82.51	89.87	93.83	95.86
Switchboard	61.80	81.24	89.41	92.84	94.70
Kids	43.79	60.38	68.46	73.53	77.37
Names	63.76	82.00	89.17	92.52	94.49
Numbers	64.80	82.19	89.49	93.17	95.12
Spelled and Spoken Words	54.33	71.03	79.90	85.21	88.83
MLTS-German	59.66	78.84	87.23	91.09	93.12
MLTS-Mandarin	53.39	73.41	83.28	88.59	91.62
MLTS-Spanish	58.65	77.71	86.34	90.43	92.71
MLTS-Japanese	57.92	78.65	87.47	91.45	93.81
MWM Diphone	49.13	72.48	84.92	90.15	93.58

closely aligned with the original TIMIT waveforms.

The average standard deviations of the alignments are 5.54 for the proposed method and 6.55 for the baseline method. The proposed method therefore has, on average, 15.4% less standard deviation on these corpora; this difference is statistically significant using a large-sample hypothesis test with a significance level of 0.05.

We also note that manual alignments of the NTIMIT corpus have notably less inter-labeler agreement (71.94% within 10 msec) than the Stories corpus (79% within 10 msec), and that inter-labeler agreement on the CTIMIT corpus is dramatically worse than on the NTIMIT corpus. These results support our claim that the addition of artificial channel distortion to the TIMIT corpus results in a decline in the presence of acoustic-phonetic cues in the new corpus (under the assumption that humans actually do use acoustic-phonetic cues when aligning speech, rather than some process closer to the motor theory). This, in turn, indicates that the relative robustness of the proposed method may be greater for non-artificial corpora, as this method relies on acoustic-phonetic information.

Table 9.4: Agreement levels for the proposed system on all corpora at several thresholds.

Corpus	Agreement within 10 msec (%)	Agreement within 20 msec (%)	Agreement within 30 msec (%)	Agreement within 40 msec (%)	Agreement within 50 msec (%)
TIMIT	80.01	92.57	96.39	98.08	98.89
FFM-TIMIT	76.26	90.62	95.32	97.38	98.40
Stories	80.24	90.24	94.31	96.22	97.40
Portland Cellular	78.13	88.89	93.27	95.57	96.94
Switchboard	71.81	85.06	90.69	93.31	94.99
Kids	59.80	74.02	79.41	83.01	86.11
Names	75.86	86.67	91.76	94.40	95.90
Numbers	74.78	85.96	90.78	93.59	95.55
Spelled and Spoken Words	69.15	80.11	85.73	89.44	91.70
MLTS-German	77.74	87.49	91.37	93.84	95.45
MLTS-Mandarin	65.22	79.33	85.87	90.34	92.98
MLTS-Spanish	72.17	82.82	87.90	90.90	93.14
MLTS-Japanese	71.73	83.41	88.72	91.69	93.44
MWM Diphone	65.73	83.88	90.92	94.42	96.34

Table 9.5: Comparison of baseline and proposed methods' agreements with manual agreements for five corpora.

Corpus (threshold)	Inter-Labeler Agreement (%)	Baseline Agreement (%)	Baseline to Manual Reduction in Error (%)	Proposed Method Agreement (%)	Proposed to Manual Reduction in Error (%)
TIMIT (20%)	93.49	89.95	35.22	92.57	12.38
Stories (10%)	79	70.77	28.16	80.24	-6.28
MLTS-Spanish (10%)	71	58.65	29.87	72.17	-4.20
MLTS-Mandarin (10%)	83	53.39	63.53	65.22	51.12
MLTS-German (10%)	79 to 81	59.66	52.90	77.74	14.65

Table 9.6: Corpora used in evaluating robustness of baseline and proposed methods (descriptions from [71, 14, 125, 52, 53]).

Corpus	Description (from documentation)
TIMIT	Recorded with Sennheiser noise-canceling, head-mounted microphone. This is the reference from which all other TIMIT variants are derived, except for FFMTIMIT.
FFMTIMIT	Recorded at same time as TIMIT, using Breul & Kjaer free-field microphone. Phonetic boundaries from TIMIT have been adjusted to correspond to FFMTIMIT data.
NTIMIT	TIMIT utterances passed through NYNEX telephone network. The waveforms have been adjusted in the time domain to align with the corresponding TIMIT waveforms. There is a maximum time discrepancy of 10 msec between the TIMIT and NTIMIT data.
CTIMIT	3367 TIMIT utterances recorded over cellular telephone channels "from a specially equipped van in a variety of driving conditions, traffic conditions, and cell sites in southern New Hampshire and Massachusetts."
HTIMIT CB1	Northern-Telecom G-type carbon-button telephone transducer (center hole membrane).
HTIMIT CB2	Northern-Telecom G-type carbon-button telephone transducer (6 hole metal).
HTIMIT CB3	Northern-Telecom G-type carbon-button telephone transducer (6 hole membrane).
HTIMIT CB4	ITT carbon-button (6 hole membrane/attached transducer).
HTIMIT EL1	Northern-Telecom Unity electret telephone transducer (3-line grill).
HTIMIT EL2	Northern-Telecom Unity Noisy-Environment electret telephone transducer (2-line grill).
HTIMIT EL3	Unknown manufacture electret (64-hole grill).
HTIMIT EL4	Radio Shack Chronophone-255 electret telephone.

9.3 Agreement for Specific Cases

We now present levels of agreement with specific types of manual boundaries for the baseline and automatic methods. The voicing detector should have a positive effect on voiced-unvoiced boundaries, the F0 information should have an effect on the boundaries between vowels and voiced consonants, the glottalization detector should have an effect on boundaries between glottalized phonemes, the measure of intensity discrimination should have an effect on silence-fricative boundaries, and the impulse detector should have an effect on the boundaries between closures and bursts. The agreement with manual alignments for the baseline and proposed methods for each of these boundary types is given in Table 9.7 based on a 20 msec threshold. It can be seen that the reduction in error for voiced-unvoiced boundaries, silence-fricative boundaries, and closure-burst boundaries is fairly high for all three corpora, ranging from 33% to 63%. The vowel-consonant boundaries have only a 4.5% reduction in error on TIMIT, but 16.6% and 20.4% for Stories and Portland Cellular, respectively. The boundaries for glottalized phonemes show a 9.3% relative *increase* in error for the proposed method on TIMIT, a 7.7% decrease in error for Stories, and a 30.4% decrease in error for Portland Cellular.

Based on these results, the voicing, intensity discrimination, and impulse-detection features seem particularly effective in the proposed method, with somewhat less effectiveness for the F0 information, and possibly negative effects for the glottalization feature. In looking at TIMIT data, we noted that sometimes the speech that is labeled with the glottalization label can have two interpretations. One interpretation is to mark glottalized speech, as characterized by irregularity in the glottal pulses. Another interpretation is indication of the silence region of a glottal stop, where the stop is characterized by a brief period of glottalization prior to or following a period of closure. In the second case, the glottalization detector output should be low and the likelihood of silence much greater during the closure region that is labeled as glottalization. This inconsistency between the acoustics and the labeled speech may quite possibly have an adverse effect on the proposed method's alignments. Because of this, the Stories and Portland Cellular corpora (which appear to have glottalization labeled more consistently) may be more indicative of the

performance of the proposed system at aligning boundaries in glottalized regions.

In order to test whether or not the neural network simply failed to properly learn the glottalization feature, we conducted a simple post-processing on the output of the proposed system's alignments. This post-processing procedure consists of the following steps:

1. Search for frames at which glottalization is detected.

2. Determine the phoneme at which glottalization occurs and the adjacent phoneme closest to the glottalization.

3. Determine where glottalization begins and ends by searching forward and backward from the glottalized frame until the detected probability of glottalization becomes zero for at least two frames.

4. If maximum glottalization in this region is strong (> 0.70), then
   (a) if the boundary between the current phoneme and adjacent phoneme is a voiced-unvoiced boundary, then move the phonetic boundary so that the glottalization is included in all of the voiced phoneme, or
   (b) if the phoneme and adjacent phoneme are both voiced, then move the phonetic boundary to the middle of the detected glottalization.

Results of this post-processing, in terms of percent agreement within 20 msec, are 91.20%, 87.35%, and 85.02% for TIMIT, Stories, and Portland Cellular, respectively. This corresponds to an 18.44%, 29.61%, and 34.83% relative increase in error over the results without post-processing, indicating that at least simple algorithms to adjust glottalization boundary alignments serve only to increase the error rate. This lends some support to the belief that the neural network has properly learned the information about glottalization.

In order to test whether the increase in error on the TIMIT corpus is due to labeling the regions of acoustic silence as well as regions of glottalization with the same label, we performed another evaluation of the glottalization boundaries. In this second evaluation, a neural network trained to recognize silence was used to check if each region labeled /q/ contained silence. Boundaries with the /q/ label were only evaluated if the /q/ region

did not contain silence. Results for this evaluation were 74.65% within 20 msec for the baseline system and 75.25% within 20 msec for the proposed system, or a 2.3% reduction in error using the proposed system. This confirms that the increase in error in the original experiment was due to the labeling discrepancy, although the reduction in error on TIMIT is not as large as on the Stories or Portland Cellular corpora.

In addition to the results for boundaries that may be affected by acoustic-level features, we present in Table 9.8 results for the 20 most common distinctive-feature combinations as evaluated on the TIMIT corpus. As expected, the proposed system shows very little improvement on approximant-vowel boundaries, because there are few cues to the exact boundary between an approximant and vowel. In addition, the proposed system has worse results than the baseline for vowel-to-silence and vowel-to-voiced fricative boundaries. We have noticed this characteristic in visual inspection of results of the proposed method, and believe that it is due to the difference in time between attenuation of the first formant and attenuation of the second formant at the end of a vowel. Often, the first formant is attenuated after the second formant, creating a region at the end of a vowel that can appear similar to a voiced closure. The proposed method may classify the attenuated-second-formant region as silence, whereas the end of a vowel is usually manually labeled at the point where *both* the first and second formants have become attenuated. Thus, the proposed method has a tendency to label the vowel-silence boundary too early. This may be resolved in the future by specifying unvoiced and voiced closures as separate categories in the proposed system. For the Place distinctive feature, boundaries involving alveolar sounds are often worse in the proposed method than in the baseline method. We hypothesize that alveolar consonants may show greater coarticulatory effects than other consonants, and that modeling alveolar sounds as context-independent may have been an over-simplification. For the Height feature, the [h3] to [max], [h4] to [h3], and [h1] to [h2] transitions are labeled with better consistency in the baseline method. It is possible that Height values for some phonemes are not categorized properly, and that adjustment of these values may improve performance in these cases.

Table 9.7: Agreement with manual labels for specific types of phonetic boundaries, for the baseline and proposed methods on three corpora. The relative reduction in error is given in the third row of each corpus.

corpus	method or reduction in error	voiced / unvoiced agreement	vowel / voiced consonant agreement	glottalized agreement	silence / fricative agreement	closure / burst agreement
TIMIT	baseline	91.06	90.54	74.57	81.38	96.65
	proposed	94.78	90.97	72.20	93.86	98.57
	reduction	**41.61**	**4.54**	**-9.32**	**67.02**	**57.31**
Stories	baseline	89.10	87.58	74.33	77.09	94.19
	proposed	92.76	89.64	76.31	86.55	97.85
	reduction	**33.58**	**16.59**	**7.71**	**41.29**	**62.99**
Portland Cellular	baseline	85.83	85.56	72.39	58.54	87.45
	proposed	93.24	88.51	80.78	75.40	94.47
	reduction	**52.29**	**20.43**	**30.39**	**40.67**	**55.94**

9.4 Influence of Acoustic Features, Distinctive Features, and Transitions

In order to estimate the relative influence of the combined acoustic-level features, the distinctive phonetic features, and the use of the proposed transition probabilities, we trained a alignment system that is identical to the baseline system, except that it uses the acoustic-level features as additional input to the neural network, and we evaluated the proposed method without the use of acoustics-dependent transition probabilities.

The baseline system with acoustic-level features has 185 inputs, 300 hidden units, and 614 outputs. It was trained on up to 8000 samples per category (for a total of nearly 2 million examples) for 45 iterations, and the best iteration was selected by evaluating forced-alignment performance on a development set.

The proposed method was evaluated without the use of acoustics-dependent transition probabilities by setting all of these probability values to 1.0; the standard duration limits were still applied.

Table 9.8: Agreement with manual labels for specific types of distinctive-feature boundaries, for the baseline and proposed methods, evaluated on the TIMIT corpus with a 20-msec threshold.

category	baseline (%)	proposed (%)	category	baseline (%)	proposed (%)	category	baseline (%)	proposed (%)
app≻vow	81.88	82.33	fnt≻alv	96.81	96.52	clo≻max	93.61	96.74
vow≻clo	92.77	95.58	alv≻clo	76.41	90.64	max≻h2	97.07	97.91
clo≻stp	96.60	98.43	clo≻alv	94.65	97.81	max≻clo	76.19	88.95
vow≻nas	95.66	96.23	alv≻fnt	97.03	97.24	max≻h3	97.87	98.39
vow≻app	74.65	75.48	fnt≻clo	91.39	95.69	h2≻max	96.34	97.06
clo≻vst	96.72	98.80	clo≻lab	94.80	96.67	max≻h4	94.03	95.34
stp≻vow	96.59	98.57	clo≻dor	96.96	98.42	h3≻max	97.13	96.97
vow≻frc	97.00	98.00	lab≻fnt	97.35	98.71	max≻max	86.65	88.06
nas≻vow	95.26	95.71	ret≻fnt	80.37	81.52	h2≻clo	93.71	94.60
frc≻vow	98.35	98.90	fnt≻fnt	87.24	87.97	max≻h1	96.07	96.68
vst≻vow	98.79	99.75	fnt≻lab	97.55	97.55	h4≻max	93.77	95.08
vow≻vfr	95.89	94.67	lab≻ret	97.65	99.13	h3≻clo	96.22	97.56
vfr≻vow	97.74	98.01	mid≻alv	97.34	97.34	h4≻h2	78.00	82.00
frc≻clo	82.22	95.34	alv≻ret	96.40	96.13	h4≻clo	77.84	88.35
nas≻clo	72.50	82.51	bck≻ret	75.95	71.87	h2≻h4	79.77	80.60
stp≻app	97.35	98.76	lat≻fnt	85.09	88.05	h1≻max	96.21	96.10
app≻clo	83.55	89.51	ret≻clo	89.67	91.25	h4≻h4	80.49	82.25
vow≻vow	71.66	74.43	alv≻mid	95.79	96.37	h4≻h3	87.54	84.93
vfr≻clo	77.00	92.81	alv≻bck	94.40	94.08	h2≻h2	70.46	75.08
app≻app	67.20	74.44	mid≻clo	96.56	96.04	h1≻h2	75.80	70.92

Results for the original baseline system, the baseline system with acoustic-level features, the proposed system without transition probabilities, and the complete proposed system are given in Table 9.9. It can be seen that the use of acoustic-level features accounts for a 9.05%, 5.69%, and 8.98% reduction in error over the baseline system on the TIMIT, Stories, and Portland Cellular corpora, respectively. The use of the complete proposed system accounts for a 28.28%, 32.22%, and 27.76% reduction in error over the system that does not use acoustics-dependent transition probabilities. The use of distinctive phonetic features alone accounts for a 13.35% and 20.70% increase in error over the baseline system that uses acoustic-level features on the TIMIT and Stories corpora, and

Table 9.9: Results for the proposed system with and without acoustic-level features and acoustics-dependent transition probabilities.

corpus	baseline system agreement	baseline plus acoustic-level features	proposed method, no transition probabilities	complete version of proposed method
TIMIT	89.95%	90.86%	89.64%	92.57%
Stories	87.35%	88.07%	85.60%	90.24%
Portland Cellular	82.51%	84.08%	84.62%	88.89%

a 3.39% decrease in error on the Portland Cellular corpus. These results assume minimal interaction between the effects of these three types of acoustic-phonetic features.

9.5 Processing Time

The proposed method operates in about 14 times real-time on a 200-MHz Pentium Pro, which is notably faster than the 150 to 400 times real-time performance of manual alignment, and about 4 times as slow as the baseline system. There are a large number of opportunities to reduce the processing time, as little effort was paid to this aspect during implementation. The first method of improving the execution time is to simply clean and optimize the code for computing the acoustic-level features. Currently, each feature uses its own module, and there is a great deal of redundancy that can be eliminated. In addition, the code for implementing the use of distinctive phonetic features was written to be flexible rather than fast; optimizing this code will also reduce the required execution time. Other methods, such as using a 10-msec frame rate, smaller neural networks, or applying a pruning threshold to the Viterbi search, will reduce the execution time, but may also reduce the accuracy of the results somewhat.

9.6 Usefulness of Improved Labels

In order to test whether the improvements in automatic alignment can have a significant effect on an application that uses phonetic alignments, we trained two HMM/ANN recognizers on the alpha-digits task. For the first recognizer, the labels for training were generated using the baseline forced-alignment system, and for the second recognizer, the labels were generated using the proposed method. In other respects the recognizers were identical; both were trained on the same speech files using as many as 8000 samples per category, both used fully-connected feed-forward networks with 300 hidden nodes, were trained for 30 iterations, and had the best iteration selected by word-level evaluation on the development set. The one minor difference between the two systems was that, because of the differences in labeling, the number of samples of each category was slightly different. As a result, the system trained on baseline alignments had three infrequent classes tied to more frequent classes (for a total of 271 output categories), whereas the system trained on the proposed system's alignments had four infrequent classes tied to more frequent classes (for a total of 270 output categories).

The test-set results were 86.88% word accuracy for the baseline system and 88.21% word accuracy for the system trained on labels from the proposed alignment method. This 10% reduction in error is significant at the 5% level using McNemar's test, with p=0.028. The results of the system trained on the proposed alignment method's labels are close to other reported results on this corpus. Hamaker et al. [61] reported results of 87.8% for a 3-state triphone HMM system built using the HTK [150] software package and results of 88.90% for a syllable-based system. We also note that our results were obtained using general-purpose alignment systems; even better results may be obtained if task-specific alignment systems are used. In addition, we look forward to training a recognition system that uses the proposed alignment method's acoustic-phonetic features, as well as the improved phonetic labels.

Chapter 10

Conclusion

10.1 Summary

In summary, the proposed method of automatic phonetic alignment using acoustic-phonetic information has significantly better agreement with manual alignments and is more robust than a state-of-the-art baseline system. On 13 corpora used for measuring agreement with manual alignments, the proposed method has an average 28% reduction in error over the baseline system, with reductions in error ranging from 20% to 41%. In some cases, the results of the proposed system are comparable to the agreement between two manual alignments. In measuring robustness, the proposed method has 14% less standard deviation in alignments when evaluated on 12 versions of the TIMIT corpus, even though in some cases the artificial means of creating the channel distortions reduced the prominence of acoustic cues that the proposed method uses.

Based on results from using acoustic-level features without distinctive phonetic features or transition information, and from using the proposed system without transition information, we conclude that the transition information provides the greatest relative improvement in performance, the acoustic-level features provide the next-greatest improvement, and the use of distinctive features may increase or decrease performance, depending on the corpus used for evaluation. The acoustic-level features that have the greatest effect on overall performance are the impulse detection, intensity discrimination, and voicing features.

In addition to these results for automatic alignment, we have shown how intensity discrimination can be used in voicing, glottalization, and impulse detection. We have

proposed and implemented a method of voicing determination that has average accuracy of 97.3% on five corpora, including microphone, telephone, and cellular speech. This represents an average 58% reduction in error over our best baseline system (which uses an improved version of the SIFT algorithm). The F0-extraction method that was developed for this thesis has accuracy of 96.88% on a corpus of read speech, with the "correct" fundamental frequency values taken from EGG data. This level of accuracy is a 45% reduction in error compared to our best baseline system. We also implemented a glottalization detection algorithm that has an average 88% accuracy on three corpora (including microphone, telephone, and cellular speech), which is a modest improvement over a baseline system. Finally, we proposed and implemented a method for the detection of burst-related impulses. This method has an equal-error rate of less than 7% on the TIMIT corpus, which is a 45% reduction in error compared to the best reported results on TIMIT.

We have also proposed and implemented a means of using acoustics-dependent transition information in the HMM framework. This method allows the use of separate recognizers for context-dependent phoneme and phonetic transition classification, performs integration of the phoneme-based and phonetic-transition-based classification results at an early stage in the alignment process, is computationally tractable, and has been shown to improve the performance of automatic alignment. One aspect of successful implementation of this method is the use of distinctive phonetic features. In our method, the distinctive phonetic features are combined using the FLMP instead of a higher-level classifier or a complex set of HMM states.

10.2 Future Work

There is still a large amount of work that can be done in developing and extending the method described in this thesis. Results of this method may be further improved by continued investigation of distinctive phonetic features and by the use of acoustic-level features that represent time-varying information (such as coarticulation or prosody). The proposed method will also have wider application if the distinctive phonetic feature set can be made more language-independent. In addition, a real-time implementation of this

forced alignment method is important for accurate, real-time synchronization of speech with automated talking agents, such as the CSLU Toolkit's Baldi.

The proposed method can also be extended to use phonological rules to adjust the phonetic transcriptions based on the acoustic evidence. Such adjustment may be useful if the transcriptions are obtained from dictionary pronunciations rather than manual determination of each word's phonetic content. For example, if the words that need to be aligned are "cost too much," and a dictionary pronunciation is used for each word, then the resulting phoneme sequence will be /k ɑ s t^h [.pau] t^h u [.pau] m ʌ ʧ/, where square brackets indicate an optional phoneme. However, it is quite likely that the two /t^h/ phonemes are merged into a single /t^h/, resulting in the sequence /k ɑ s [.pau] t^h u [.pau] m ʌ ʧ/. Such rules can be used to obtain a phonetic sequence that is more likely to match the speech signal. Forced alignment can be done with and without such rules, and the phonetic sequence with the best score can be chosen as the correct sequence.

The alignment systems that have been developed for this thesis are general-purpose systems. In the same way that task-specific or speaker-specific recognizers usually outperform general-purpose recognizers on that task or speaker, it is probable that a task-specific or speaker-specific alignment system will have better results than the systems implemented here on a given task or speaker. For example, in aligning the high-quality speech of a single speaker to be used in text-to-speech synthesis, we noted that the automatically-aligned boundary between a vowel and an /n/ tended to occur later than expected. It was found that this speaker strongly articulated the /n/ and that formants for this phoneme in the 2000-Hz region were stronger than average. This caused the distinctive-feature networks to return values more appropriate for a vowel than for a consonant during the /n/, thereby resulting in a poor boundary placement. There are two approaches to a solution to such problems. First, there is a data-driven approach, in which a new alignment system is trained on the given task or speaker. This assumes that there is at least some amount of data for this task or speaker that have been accurately aligned. Second, there is a knowledge-based approach, in which the errors are analyzed, and changes are made to the existing system to address the cause of these errors. In the case of the data mentioned above, the description of an /n/ in terms of distinctive phonetic features could be modified

to include properties of the observed /n/, such as a Height value of [h4] instead of [max]. The data-driven and knowledge-based approaches can, of course, be combined.

Finally, we are excited about the prospects of using the methods described in this thesis for training speech recognition systems. Although the acoustic-level features can be easily integrated into recognition systems, the Viterbi search that uses the transition probabilities and distinctive phonetic features was written specifically for forced alignment. Some additional effort will be required to extend the current implementation to be capable of word recognition. We hope to begin work on these improvements shortly.

Bibliography

[1] ALLEN, J. B. How Do Humans Process and Recognize Speech? *IEEE Transactions on Speech and Audio Processing, 2*, 4 (October 1990), 567–577.

[2] ANDERSSON, Å. A Method for Coarse Speech-to-Text Alignment. *Technical Report CTH-TE-39*, Chalmers University of Technology, Department of Applied Electronics, Göteborg, Sweden, February 1996.

[3] ANGELINI, B., BRUGNARA, F., FALAVIGNA, D., GUILIANI, D., GRETTER, R., AND OMOLOGO, M. Automatic Segmentation and Labeling of English and Italian Speech Databases. In *Proceedings of Eurospeech '93* (Berlin, Germany, September 1993), pp. 653–656.

[4] ATAL, B. S., AND RABINER, L. R. A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Applications to Speech Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP-24*, 3 (June 1976), 201–212.

[5] BARNARD, E., COLE, R. A., VEA, M. P., AND ALLEVA, F. A. Pitch Detection with a Neural-Net Classifier. *IEEE Transactions on Signal Processing, 39*, 2 (February 1991), 298–307.

[6] BELL, C. G., FUJISAKI, H., HEINZ, J. M., STEVENS, K. N., AND HOUSE, A. S. Reduction of Speech by Analysis-by-Synthesis Techniques. *Journal of the Acoustical Society of America, 33*, 12 (1961), 1725–1736.

[7] BENGIO, Y. Markovian Models for Sequential Data. *Neural Computing Surveys, 2* (1999), 129–162.

[8] BENGIO, Y., AND FRASCONI, P. An Input-Output HMM Architecture. In *Advances in Neural Information Processing Systems (NIPS 7)*, G. Tesauro, D. S. Touretzky, and T. K. Leen, eds. MIT Press, 1995.

[9] BLOMBERG, M., AND CARLSON, R. Labelling of Speech Given its Text Representation. In *Proceedings of Eurospeech '93* (Berlin, Germany, September 1993), pp. 1775–1778.

[10] BOURLARD, H. Towards Increasing Speech Recognition Error Rates. In *Proceedings of Eurospeech '95* (Madrid, Spain, September 1995), pp. 883–894.

[11] BOURLARD, H., MORGAN, N., AND RENALS, S. Neural Nets and Hidden Markov Models: Review and Generalizations. *Speech Communication, 11* (1992), 237–246.

[12] BOURLARD, H. A., AND MORGAN, N. *Connectionist Speech Recognition: A Hybrid Approach.* Kluwer Academic Publishers, 1994.

[13] BREGMAN, A. S. *Auditory Scene Analysis.* The MIT Press, 1990.

[14] BROWN, K. L., AND GEORGE, E. B. CTIMIT: A Speech Corpus for the Cellular Environment with Applications to Automatic Speech Recognition. In *Proceedings of ICASSP '95* (Detroit, MI, May 1995), pp. 105–108.

[15] BRUGNARA, F., FALAVIGNA, D., AND OMOLOGO, M. A HMM-Based System for Automatic Segmentation and Labeling of Speech. In *Proceedings of ICSLP '92* (Banff, Alberta, Canada, October 1992), pp. 803–806.

[16] BRUGNARA, F., FALAVIGNA, D., AND OMOLOGO, M. Automatic Segmentation and Labeling of Speech Based on Hidden Markov Models. *Speech Communication, 12*, 4 (1993), 357–370.

[17] BURGES, C. J. C. *A Tutorial on Support Vector Machines for Pattern Recognition.* Kluwer Academic Publishers, Boston, MA, 1998. http://svm.research.bell-labs.com/SVMrefs.html. Date viewed: April 18, 2000.

[18] BURSHTEIN, D. Robust Parametric Modeling of Durations in Hidden Markov Models. In *Proceedings of ICASSP '95* (Detroit, MI, May 1995), pp. 548–551.

[19] CAMPBELL, N. Autolabelling Japanese ToBI. In *Proceedings of ICSLP '96* (Philadelphia, PA, October 1996), pp. 2399–2402.

[20] CHAMERLAIN, R. M., AND BRIDLE, J. S. ZIP: A Dynamic Programming Algorithm for Time-Aligning Two Indefinitely Long Utterances. In *Proceedings of ICASSP '83* (Boston, MA, 1983), pp. 816–819.

[21] CHANG, J. W., AND GLASS, J. R. Segmentation and Modeling in Segment-Based Recognition. In *Proceedings of Eurospeech '97* (Rhodes, Greece, September 1997), pp. 1199–1202.

[22] CHUNG, G., AND SENEFF, S. Hierarchical Duration Modelling for Speech Recognition Using the ANGIE Framework. In *Proceedings of Eurospeech '97* (Rhodes, Greece, September 1997), pp. 1475–1478.

[23] COHEN, M. M., AND MASSARO, D. W. Modeling Coarticulation in Synthetic Visual Speech. In *Models and Techniques in Computer Animation*, N. M. Thalmann and D. Thalmann, eds. Springer-Verlag, Tokyo, Japan, 1993.

[24] COLE, R., FANTY, M., MUTHUSAMY, Y., AND GOPALAKRISHNAN, M. Speaker-Independent Recognition of Spoken English Letters. In *Proceedings of the International Joint Conference on Neural Networks* (San Diego, CA, June 1990), vol. 2, pp. 45–51.

[25] COLE, R., OSHIKA, B. T., NOEL, M., LANDER, T., AND FANTY, M. Labeler Agreement in Phonetic Labeling of Continuous Speech. In *Proceedings of ICSLP '94* (Yokohama, Japan, September 1994), pp. 2131–2134.

[26] COLE, R. A. Spoken Language Technology. In *More Than Screen Deep*, N. R. Council, ed. National Academy Press, Washington, D.C., 1997.

[27] COLE, R. A., NOEL, M., LANDER, T., AND DURHAM, T. New Telephone Speech Corpora at CSLU. In *Proceedings of Eurospeech '95* (Madrid, Spain, September 1995), pp. 821–824.

[28] COLE, R. A., AND SCOTT, B. Toward a Theory of Speech Perception. *Psychological Review, 81*, 4 (July 1974), 348–374.

[29] COLE, R. A., ZUE, V. W., AND REDDY, D. R. Speech as Patterns on Paper. In *Perception and Production of Fluent Speech*, R. A. Cole, ed. Lawrence Erlbaum Associates, Hillsdale, NY, 1980.

[30] COOKE, M. P., AND BROWN, G. J. Computational Auditory Scene Analysis: Exploiting Principles of Perceived Continuity. *Speech Communication, 13*, 3 (1993), 391–399.

[31] COSI, P. SLAM: Segmentation and Labelling Automatic Module. In *Proceedings of Eurospeech '93* (Berlin, Germany, September 1993), vol. 1, pp. 88–91.

[32] COSI, P., FALAVIGNA, D., AND OMOLOGO, M. A Preliminary Statistical Evaluation of Manual and Automatic Segmentation Discrepancies. In *Proceedings of Eurospeech '91* (Genova, Italy, September 1991), pp. 693–696.

[33] COX, S., BRADY, R., AND JACKSON, P. Techniques for Accurate Automatic Annotation of Speech Waveforms. In *Proceedings of ICSLP '98* (Sydney, Australia, December 1998), vol. 5, pp. 1947–1950.

[34] CRAVERO, M., PIERACCINI, R., AND RAINERI, F. Definition and Evaluation of Phonetic Units for Speech Recognition by Hidden Markov Models. In *Proceedings of ICASSP '86* (Tokyo, Japan, April 1986), pp. 2235–2238.

[35] DALSGAARD, P. Phoneme Label Alignment Using Acoustic-Phonetic Features and Gaussian Probability Density Functions. *Computer Speech and Language, 6*, 4 (1992), 303–329.

[36] DALSGAARD, P., ANDERSEN, O., AND BARRY, W. Multi-Lingual Label Alignment Using Acoustic-Phonetic Features Derived by Neural-Network Technique. In *Proceedings of ICASSP '91* (Toronto, Canada, May 1991), pp. 197–200.

[37] DALSGAARD, P., ANDERSEN, O., BARRY, W., AND JØRGENSEN, R. On the Use of Acoustic-Phonetic Features in Interactive Labelling of Multi-Lingual Speech Corpora. In *Proceedings of ICASSP '92* (San Francisco, CA, March 1992), vol. 1, pp. 549–552.

[38] DANHAUER, J. L., AND SINGH, S. *Multidimensional Speech Perception by the Hearing Impaired.* University Park Press, Baltimore, MD, 1975, pp. 38–44.

[39] DAVIS, S., AND MERMELSTEIN, P. Comparison of Parametric Representations for Monosyllabic Word Recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-28*, 4 (1980), 357–366.

[40] DELATTRE, P. C., LIBERMAN, A. M., AND COOPER, F. S. Acoustic Loci and Transitional Cues for Consonants. *Journal of the Acoustical Society of America, 27*, 4 (1955), 769–773.

[41] DENG, L. Speech Recognition Using Autosegmental Representation of Phonological Units with Interface to the Trended HMM. *Speech Communication, 23*, 3 (1997), 211–222.

[42] DENG, L., AND SUN, D. X. A Statistical Approach to Automatic Speech Recognition Using the Atomic Speech Units Constructed from Overlapping Articulatory Features. *Journal of the Acoustical Society of America, 95*, 5, Pt. 1 (May 1994).

[43] DUTOIT, T., PAGEL, V., PIERRET, N., BATAILLE, F., AND DER VRECKEN, O. V. The MBROLA project: Towards a Set of High Quality Speech Synthesizers Free of Use for Non Commercial Purposes. In *Proceedings of ICSLP '96* (Philadelphia, PA, October 1996), vol. 3, pp. 1393–1396.

[44] ERLER, K., AND FREEMAN, G. H. An HMM-Based Speech Recognizer Using Overlapping Articulatory Features. *Journal of the Acoustical Society of America, 100*, 4 (October 1996), 2500–2513.

[45] FALAVIGNA, D., AND OMOLOGO, M. A DTW-Based Approach to the Automatic Labeling of Speech According to the Phonetic Transcription. In *Proceedings of the European Signal Processing Conference* (Barcelona, Spain, 1990), pp. 1139–1142.

[46] FANT, G. *Acoustic Theory of Speech Production with Calculations Based on X-Ray Studies of Russian Articulation.* Mouton, The Hague, 1970.

[47] FANT, G., LILJENCRANTS, J., AND LIN, Q. A Four-Parameter Model of Glottal Flow. *Quarterly Progress and Status Report 4*, The Royal Institute of Technology (KTH), Department of Speech, Music, and Hearing, Speech Transmission Laboratory, Stockholm, Sweden, 1985.

[48] FUJISAKI, H., AND LJUNGQVIST, M. Proposal and Evaluation of Models of the Glottal Source Waveform. In *Proceedings of ICASSP '86* (Tokyo, Japan, April 1986), pp. 1605–1608.

[49] FUJISAKI, H., AND TANABE, Y. A Time-Domain Technique for Pitch Extraction of Speech. *Annual Report of the Engineering Research Institute, Faculty of Engineering, University of Tokyo, 31* (September 1972), 213–220.

[50] FUJISAKI, H., AND TANABE, Y. A Time-Domain Technique for Pitch Extraction of Speech. *Journal of the Acoustical Society of Japan, 29*, 7 (July 1973), 418–419.

[51] FURUI, S. On the Role of Spectral Transitions for Speech Perception. *Journal of the Acoustical Society of America, 80*, 4 (1986), 1016–1025.

[52] GAROFOLO, J. S., LAMEL, L. F., FISHER, W. M., FISCUS, J. G., PALLETT, D. S., AND DAHLGREN, N. L. DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM, 1990. National Institute of Standards and Technology, NTIS Order No. PB91-505065.

[53] GAROFOLO, J. S., LAMEL, L. F., FISHER, W. M., FISCUS, J. G., PALLETT, D. S., AND DAHLGREN, N. L. FFM TIMIT Acoustic-Phonetic Continuous Speech Corpus Secondary (Far Field) Microphone Recordings CD-ROM, 1993. National Institute of Standards and Technology, NTIS Order NO. PB95-504569.

[54] GHOLAMPOUR, I., AND NAYEBI, K. A New Fast Algorithm for Automatic Segmentation of Continuous Speech. In *Proceedings of ICSLP '98* (Sydney, Australia, December 1998), vol. 4, pp. 1555–1558.

[55] GILLICK, L., AND COX, S. J. Some Statistical Issues in the Comparison of Speech Recognition Algorithms. In *Proceedings of ICASSP '89* (Glasgow, Scotland, May 1989), pp. 532–535.

[56] GLASS, J., CHANG, J., AND MCCANDLESS, M. A Probabilistic Framework for Feature-based Speech Recognition. In *Proceedings of ICSLP '96* (Philadelphia, PA, October 1996), vol. 4, pp. 2277–2280.

[57] GLASS, J. R., AND ZUE, V. W. Multi-Level Acoustic Segmentation of Continuous Speech. In *Proceedings of ICASSP '88* (New York, NY, 1988), pp. 429–432.

[58] GONG, Y., AND HATON, J.-P. Iterative Transformation and Alignment for Speech Labeling. In *Proceedings of Eurospeech '93* (Berlin, Germany, September 1993), pp. 1759–1762.

[59] GREENBERG, S. Understanding Speech Understanding: Towards a Unified Theory of Speech Perception. In *Proceedings of the ESCA Tutorial and Advanced Research Workshop on the Auditory Basis of Speech Perception* (Keele, England, July 1996), pp. 1–8.

[60] GREENBERG, S., AND KINGSBURY, B. The Modulation Spectrogram: In Pursuit of an Invariant Representation of Speech. In *Proceedings of ICASSP '97* (Munich, Germany, April 1997), vol. 3, pp. 1647–1650.

[61] HAMAKER, J., GANAPATHIRAJU, A., PICONE, J., AND GODFREY, J. J. Advances in Alphadigit Recognition Using Syllables. In *Proceedings of ICASSP '98* (Seattle, WA, May 1998), vol. 1, pp. 421–424.

[62] HARRIS, J. D., AND NELSON, D. Glottal Pulse Alignment in Voiced Speech for Pitch Determination. In *Proceedings of ICASSP '93* (Minneapolis, MN, April 1993), vol. 2, pp. 519–522.

[63] HERMANSKY, H. Perceptual Linear Predictive (PLP) Analysis of Speech. *Journal of the Acoustical Society of America, 87*, 4 (April 1990), 1738–1752.

[64] HIERONYMUS, J. L. ASCII Phonetic Symbols for the World's Languages: Worldbet. *Technical Report*, AT&T Bell Laboratories, Murray Hill, NJ, February 1995.

[65] HOSOM, J.-P. Investigating Additional Features for Improved Speech Recognition. *Research proficiency examination written report*, Oregon Graduate Institute of Science and Technology, Center for Spoken Language Understanding, Beaverton, OR, May 1996.

[66] HOSOM, J.-P. A Comparison of Speech Recognizers Created Using Manually-Aligned and Automatically-Aligned Training Data. *Technical Report CSE-00-002*, Oregon Graduate Institute of Science and Technology, Center for Spoken Language Understanding, Beaverton, OR, January 2000.

[67] HOSOM, J.-P., AND COLE, R. A. A Diphone-Based Digit Recognition System. In *Proceedings of ICASSP '97* (Munich, Germany, April 1997), vol. 4, pp. 3369–3372.

[68] HOSOM, J.-P., COSI, P., AND COLE, R. A. Evaluation and Integration of Neural-Network Training Techniques for Continuous Digit Recognition. In *Proceedings of ICSLP '98* (Sydney, Australia, December 1998), vol. 3, pp. 731–734.

[69] HÜBENER, K., AND CARSON-BERNDSEN, J. Phoneme Recognition Using Acoustic Events. In *Proceedings of ICSLP '94* (Yokohama, Japan, September 1994), pp. 1919–1922.

[70] HUBER, D. Perception of Aperiodic Speech Signals. In *Proceedings of ICSLP '92* (Banff, Alberta, Canada, October 1992), pp. 503–506.

[71] JANKOWSKI, C., KALYANSWAMY, A., BASSON, S., AND SPITZ, J. NTIMIT: A Phonetically Balanced, Continuous Speech, Telephone Bandwidth Speech Database. In *Proceedings of ICASSP '90* (Albuquerque, NM, April 1990), pp. 109–112.

[72] KANEDERA, N., ARAI, T., HERMANSKY, H., AND PAVEL, M. On the Importance of Various Modulation Frequencies for Speech Recognition. In *Proceedings of Eurospeech '97* (Rhodes, Greece, September 1997), pp. 1079–1082.

[73] KARJALAINEN, M., ALTOSAAR, T., AND HUTTUNEN, M. An Efficient Labeling Tool for the Quicksig Speech Database. In *Proceedings of ICSLP '98* (Sydney, Australia, December 1998), vol. 4, pp. 1535–1538.

[74] KENT, R. D., AND MINIFIE, F. D. Coarticulation in Recent Speech Production Models. *Journal of Phonetics, 5*, 2 (1977), 115–133.

[75] KIPP, A., WESENICK, M.-B., AND SCHIEL, F. Automatic Detection and Segmentation of Pronunciation Variants in German Speech Corpora. In *Proceedings of ICSLP '96* (Philadelphia, PA, October 1996), pp. 106–109.

[76] KIRCHHOFF, K. Syllable-level Desynchronisation of Phonetic Features for Speech Recognition. In *Proceedings of ICSLP '96* (Philadelphia, PA, October 1996), vol. 4, pp. 2274–2276.

[77] KIRCHHOFF, K. Combining Articulatory and Acoustic Information for Speech Recognition in Noisy and Reverberant Environments. In *Proceedings of ICSLP '98* (Sydney, Australia, December 1998), vol. 3, pp. 891–894.

[78] KITAZAWA, S., AND SERIZAWA, M. An Artificial Neural Network for the Burst Point Detection. In *Proceedings of ICSLP '90* (Kobe, Japan, November 1990), pp. 1069–1072.

[79] KLATT, D. H. Linguistic Uses of Segmental Duration in English: Acoustic and Perceptual Evidence. *Journal of the Acoustical Society of America, 59*, 5 (May 1976), 1208–1221.

[80] KLATT, D. H. Review of Text-to-Speech Conversion for English. *Journal of the Acoustical Society of America, 82*, 3 (1987), 737–793.

[81] KLATT, D. H., AND KLATT, L. Analysis, Synthesis, and Perception of Voice Quality Variations Among Female and Male Talkers. *Journal of the Acoustical Society of America, 87*, 2 (1990), 820–857.

[82] KOREMAN, J., ANDREEVA, B., AND BARRY, W. J. Do Phonetic Features Help to Improve Consonant Identification in ASR? In *Proceedings of ICSLP '98* (Sydney, Australia, December 1998), vol. 3, pp. 1035–1038.

[83] KVALE, K. On the Connection Between Manual Segmentation Conventions and "Errors" Made by Automatic Segmentation. In *Proceedings of ICSLP '94* (Yokohama, Japan, September 1994), vol. 3, pp. 1667–1670.

[84] LADEFOGED, P. *A Course in Phonetics.* Harcourt Brace College Publishers, Fort Worth, TX, 1993, p. 275.

[85] LADEFOGED, P. *A Course in Phonetics.* Harcourt Brace College Publishers, Fort Worth, TX, 1993, p. 42.

[86] LADEFOGED, P. *A Course in Phonetics.* Harcourt Brace College Publishers, Fort Worth, TX, 1993, pp. 168–169.

[87] LAMEL, L., AND GAUVIN, J. L. High-Performance Speaker-Independent Phone Recognition Using CDHMM. In *Proceedings of Eurospeech '93* (Berlin, Germany, September 1993), pp. 121–124.

[88] LEE, S. C., AND GLASS, J. R. Real-Time Probabilistic Segmentation for Segment-Based Speech Recognition. In *Proceedings of ICSLP '98* (Sydney, Australia, December 1998), vol. 5, pp. 1803–1806.

[89] LEUNG, H. C., AND ZUE, V. W. A Procedure for Automatic Alignment of Phonetic Transcriptions with Continuous Speech. In *Proceedings of ICASSP '84* (San Diego, California, 1984), pp. 2.7.1–2.7.4.

[90] LIBERMAN, A. M., AND MATTINGLY, I. G. The Motor Theory of Speech Perception Revised. *Cognition, 21*, 1 (1985), 1–36.

[91] LIU, S. A. Noise Effects on Landmark Detection in a Speech Recognition System. In *Proceedings of Eurospeech '95* (Madrid, Spain, September 1995), pp. 1567–1570.

[92] LJOLJE, A., HIRSCHBERG, J., AND VAN SANTEN, J. P. H. Automatic Speech Segmentation for Concatenative Inventory Selection. In *Progress in Speech Synthesis*, J. P. H. V. Santen, R. W. Sproat, J. Olive, and J. Hirschberg, eds. Springer-Verlag, New York, 1997.

[93] LJOLJE, A., AND RILEY, M. D. Automatic Segmentation and Labeling of Speech. In *Proceedings of ICASSP '91* (Toronto, Canada, May 1991), pp. 473–476.

[94] LOIZOU, P. C., AND SPANIAS, A. S. High-Performance Alphabet Recognition. *IEEE Transactions on Speech and Audio Processing, 4*, 6 (November 1996), 430–444.

[95] LUCE, R. D. *Individual Choice Behavior: A Theoretical Analysis.* J. Wiley and Sons, New York, NY, 1959.

[96] MALFRÈRE, F., DEROO, O., AND DUTOIT, T. Phonetic Alignment: Speech Synthesis vs. Hybrid HMM/ANN. In *Proceedings of ICSLP '98* (Sydney, Australia, December 1998), vol. 4, pp. 1571–1574.

[97] MANCERON, F., AND LIENARD, J. S. Impulse Analysis of Speech: Spotting and Preclassifying the Impulses in the Speech Wave. In *Proceedings of ICASSP '82* (New York, NY, 1982), pp. 1569–1572.

[98] MARASEK, K. Automatic Classification of Voice Quality Using the EGG Signal. *Technical Report*, Stuttgart University, Experimental Phonetics Group, Stuttgart, Germany, 1997. http://www.ims.uni-stuttgart.de/phonetik/EGG/frmst2.htm. Date viewed: April 18, 2000.

[99] MARKEL, J. D. The SIFT Algorithm for Fundamental Frequency Estimation. *IEEE Transactions on Audio and Electroacoustics, AU-20*, 5 (December 1972), 367–377.

[100] MASSARO, D. W., AND FRIEDMAN, D. Models of Integration Given Multiple Sources of Information. *Psychological Review, 97*, 2 (1990), 225–252.

[101] McClelland, J. L., and Elman, J. L. The TRACE Model of Speech Perception. *Cognitive Psychology, 18*, 1 (1986), 1–86.

[102] Moore, B. C. J. *An Introduction to the Psychology of Hearing.* Academic Press, San Diego, CA, 1997, pp. 63–65.

[103] Moore, B. C. J. *An Introduction to the Psychology of Hearing.* Academic Press, San Diego, CA, 1997, p. 361.

[104] Moore, B. C. J. *An Introduction to the Psychology of Hearing.* Academic Press, San Diego, CA, 1997, p. 212.

[105] Moore, R. K. Whither a Theory of Speech Pattern Processing? In *Proceedings of Eurospeech '93* (Berlin, Germany, September 1993), pp. 43–47.

[106] Morgan, N., Bourlard, H., Greenberg, S., and Hermansky, H. Stochastic Perceptual Auditory-Event-Based Models for Speech Recognition. In *Proceedings of ICSLP '94* (Yokohama, Japan, September 1994), pp. 1943–1946.

[107] Morris, A. C., and Pardo, J. M. Phoneme Transition Detection and Broad Classification Using a Simple Model Based on the Function of Onset Detector Cells Found in the Cochlear Nucleus. In *Proceedings of Eurospeech '95* (Madrid, Spain, September 1995), pp. 115–118.

[108] Niyogi, P., Burges, C., and Ramesh, P. Distinctive Feature Detection Using Support Vector Machines. In *Proceedings of ICASSP '99* (Phoenix, AZ, March 1998), pp. 425–428.

[109] Niyogi, P., Mitra, P., and Sondhi, M. M. A Detection Framework for Locating Phonetic Events. In *Proceedings of ICSLP '98* (Sydney, Australia, December 1998), vol. 3, pp. 1067–1070.

[110] Noll, A. M. Cepstrum Pitch Detection. *Journal of the Acoustical Society of America, 41*, 2 (February 1967), 293–309.

[111] Oden, G. G., and Massaro, D. W. Integration of Featural Information in Speech Perception. *Psychological Review, 85*, 3 (1978), 172–191.

[112] Öhman, S. E. G. Coarticulation in VCV Utterances: Spectrographic Measurements. *Journal of the Acoustical Society of America, 39* (1966), 151–168.

[113] Pauws, S., Kamp, Y., and Willems, L. A Hierarchical Method of Automatic Speech Segmentation for Synthesis Applications. *Speech Communication, 19*, 4 (1996), 207–220.

[114] PELLOM, B. L. *Enhancement, Segmentation, and Synthesis of Speech with Application to Robust Speaker Recognition.* PhD thesis, Duke University, Durham, North Carolina, 1998.

[115] PELLOM, B. L., AND HANSEN, J. H. L. Automatic Segmentation and Labeling of Speech Recorded in Unknown Noisy Channel Environments. In *Proceedings of the 1997 ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels* (1997), pp. 167–170.

[116] PETEK, B., ANDERSEN, O., AND DALSGAARD, P. On the Robust Automatic Segmentation of Spontaneous Speech. In *Proceedings of ICSLP '96* (Philadelphia, PA, October 1996), pp. 913–916.

[117] PICKETT, J. M. *The Sounds of Speech Communication.* University Park Press, Baltimore, MD, 1980, pp. 202–206.

[118] RABINER, L., AND JUANG, B. *Fundamentals of Speech Recognition.* Prentice Hall, Englewood Cliffs, NJ, 1993, p. 19.

[119] RABINER, L., AND JUANG, B. *Fundamentals of Speech Recognition.* Prentice Hall, Englewood Cliffs, NJ, 1993, pp. 321–389.

[120] RABINER, L., AND JUANG, B. *Fundamentals of Speech Recognition.* Prentice Hall, Englewood Cliffs, NJ, 1993, p. 370.

[121] RABINER, L., AND JUANG, B. *Fundamentals of Speech Recognition.* Prentice Hall, Englewood Cliffs, NJ, 1993, pp. 358–362.

[122] RABINER, L., AND JUANG, B. *Fundamentals of Speech Recognition.* Prentice Hall, Englewood Cliffs, NJ, 1993, pp. 342–344.

[123] RAPP, S. Automatic Phonemic Transcription and Linguistic Annotation from Known Text with Hidden Markov Models / An Aligner for German. In *Proceedings of ELSNET Goes East and IMACS Workshop* (Moscow, Russia, 1995). URL http://www.ims.uni-stuttgart.de/~rapp/. Date viewed: April 18, 2000.

[124] RENALS, S., MORGAN, N., BOURLARD, H., COHEN, M., AND FRANCO, H. Connectionist Probability Estimators in HMM Speech Recognition. *IEEE Transactions on Speech and Audio Processing, 2*, 1 (January 1994), 161–174.

[125] REYNOLDS, D. A. HTIMIT and LLHDB: Speech Corpora for the Study of Handset Transducer Effects. In *Proceedings of ICASSP '97* (Munich, Germany, April 1997), pp. 1535–1538.

[126] RICHARD, M. D., AND LIPPMANN, R. P. Neural Network Classifiers Estimate Bayesian *a posteriori* Probabilities. *Neural Computation, 3*, 4 (1991), 461–483.

[127] RIIS, S. K., AND KROGH, A. Hidden Neural Networks: A Framework for HMM/NN Hybrids. In *Proceedings of ICASSP '97* (Munich, Germany, April 1997), vol. 4, pp. 3233–3236.

[128] ROBINSON, A. An Application of Recurrent Neural Nets to Phone Probability Estimation. *Transactions on Neural Networks, 5*, 2 (March 1994), 298–305.

[129] SAITO, T. On the Use of F0 Features in Automatic Segmentation for Speech Synthesis. In *Proceedings of ICSLP '98* (Sydney, Australia, December 1998), vol. 7, pp. 2839–2842.

[130] SCHMID, P. *Explicit N-Best Formant Features for Segment-Based Speech Recognition.* PhD thesis, Oregon Graduate Institute of Science and Technology, Beaverton, Oregon, October 1996.

[131] SCHMIDBAUER, O. Robust Statistical Modeling of Systematic Variabilities in Continuous Speech Incorporating Acoustic-Articulatory Relations. In *Proceedings of ICASSP '89* (Glasgow, Scotland, May 1989), pp. 616–619.

[132] SHOBAKI, K., COLE, R., AND COLE, D. *The OGI Kids' Speech Corpus.* URL=http://cslu.cse.ogi.edu/corpora/kids/. Date viewed: April 18, 2000.

[133] STEVENS, K. N., AND BLUMSTEIN, S. E. Invariant Cues for Place of Articulation in Stop Consonants. *Journal of the Acoustical Society of America, 64*, 5 (1978), 1358–1368.

[134] STÖBER, K., AND HESS, W. Additional Use of Phoneme Duration Hypotheses in Automatic Speech Segmentation. In *Proceedings of ICSLP '98* (Sydney, Australia, December 1998), vol. 4, pp. 1595–1598.

[135] SUH, Y., HWANG, K., KWON, O., AND PARK, J. Improving Speech Recognizer by Broader Acoustic-Phonetic Group Classification. In *Proceedings of ICSLP '98* (Sydney, Australia, December 1998), vol. 3, pp. 1107–1110.

[136] SVENDSEN, T., AND KVALE, K. Automatic Alignment of Phonemic Labels with Continuous Speech. In *Proceedings of ICSLP '90* (Kobe, Japan, November 1990), pp. 997–1000.

[137] SVENDSEN, T., AND SOONG, F. K. On the Automatic Segmentation of Speech Signals. In *Proceedings of ICASSP '87* (Dallas, TX, April 1987), pp. 77–80.

[138] TIBREWALA, S., AND HERMANSKY, H. Sub-Band Based Recognition of Noisy Speech. In *Proceedings of ICASSP '97* (Munich, Germany, April 1997), vol. 2, pp. 1255–1258.

[139] TORKKOLA, K. Automatic Alignment of Speech with Phonetic Transcriptions in Real Time. In *Proceedings of ICASSP '88* (New York, NY, 1988), pp. 611–614.

[140] VAN SANTEN, J. P. H. Contextual Effects on Vowel Duration. *Speech Communication, 11*, 6 (1992), 513–546.

[141] VAN SANTEN, J. P. H., AND SPROAT, R. W. High-Accuracy Automatic Segmentation. In *Proceedings of Eurospeech '99* (Budapest, Hungary, September 1999), vol. 6, pp. 2809–2812.

[142] VORSTERMANS, A., MARTENS, J.-P., AND COILE, B. V. Automatic Segmentation and Labelling of Multi-Lingual Speech Data. *Speech Communication, 19*, 4 (1996), 271–293.

[143] WAGNER, M. Automatic Labelling of Continuous Speech with a Given Phonetic Transcription Using Dynamic Programming Algorithms. In *Proceedings of ICASSP '81* (Atlanta, GA, 1981), pp. 1156–1159.

[144] WANG, H. D., BAILLY, G., AND TUFFELLI, D. Automatic Segmentation and Alignment of Continuous Speech Based on Temporal Decomposition Model. In *Proceedings of ICSLP '96* (Philadelphia, PA, October 1996), pp. 457–460.

[145] WEI, W., AND VAN VUUREN, S. Improved Neural Network Training of Inter-Word Context Units for Connected Digit Recognition. In *Proceedings of ICASSP '98* (Seattle, WA, May 1998), vol. 1, pp. 497–500.

[146] WESENICK, M.-B., September 1999. e-mail communication.

[147] WESENICK, M.-B., AND KIPP, A. Estimating the Quality of Phonetic Transcriptions and Segmentations of Speech Signals. In *Proceedings of ICSLP '96* (Philadelphia, PA, October 1996), pp. 129–132.

[148] WHEATLEY, B., DODDINGTON, G., HEMPHILL, C., GODFREY, J., HOLLIMAN, E., MCDANIEL, J., AND FISHER, D. Robust Automatic Time Alignment of Orthographic Transcriptions with Unconstrained Speech. In *Proceedings of ICASSP '92* (San Francisco, CA, March 1992), vol. 1, pp. 533–536.

[149] WIGHTMAN, C. W., AND TALKIN, D. T. The Aligner: Text-to-Speech Alignment Using Markov Models. In *Progress in Speech Synthesis*, J. P. H. V. Santen, R. W. Sproat, J. Olive, and J. Hirschberg, eds. Springer-Verlag, New York, 1997.

[150] WOODLAND, P. C., LEGGETTER, C. J., ODELL, J. J., VALTCHEV, V., AND YOUNG, S. The 1994 HTK Large Vocabulary Speech Recognition System. In *Proceedings of ICASSP '95* (Detroit, MI, May 1995), pp. 73–76.

[151] WU, S.-L. Properties of Stochastic Perceptual Auditory-Event-Based Models for Automatic Speech Recognition. *Technical Report TR-95-023*, ICSI, University of California at Berkeley, Berkeley, CA, May 1995.

[152] WU, S.-L., SHIRE, M. L., GREENBERG, S., AND MORGAN, N. Integrating Syllable Boundary Information into Speech Recognition. In *Proceedings of ICASSP '97* (Munich, Germany, April 1997), pp. 987–990.

[153] YAN, Y., FANTY, M., AND COLE, R. Speech Recognition Using Neural Networks with Forward-Backward Probability Generated Targets. In *Proceedings of ICASSP '97* (Munich, Germany, April 1997), vol. 4, pp. 3241–3244.

Appendix A

Stochastic Frame-Based Speech Recognition

A.1 HMM Framework

The basic framework for HMM speech recognition is illustrated in Figures A.1, A.2, and A.3. As mentioned in Section 2.3.2, a typical HMM system works by dividing the speech into short frames, where each frame corresponds to a state in a state sequence. Phonetic-based recognition is performed on each frame, and the most likely word-level path through the state sequence is computed using a dynamic-programming search called a Viterbi search. The notation and explanations in this chapter are based on a book by Rabiner and Juang [119] and papers by Bourlard, Renals, Morgan, and others [124, 11, 12].

In Figure A.1, an HMM for a simple two-word vocabulary is shown. Each state is associated with phonetic likelihoods, and states are connected by unidirectional links. In this figure, each state is marked by a phonetic symbol; these symbols can be thought of as the most likely phonetic observation that will occur in that state. A series of connected states forms a word, and word-ending states can be connected to word-beginning states in order to create a continuous-speech recognizer. Each arc between two states has an associated probability of transitioning from its "previous" state to its "current" state. These probabilities are referred to as "transition probabilities," and are denoted a_{ij} (where i is the previous state and j is the current state). In most HMM systems, each state has a self-loop (a_{ii}), which allows the HMM to remain in the same state for more than one time frame.

In a given state, the probabilities of observing each phoneme in that state are referred to as "observation probabilities," and are denoted $b_j(\mathbf{o}_t)$ or $b_j(k)$, where j is the given state, $\mathbf{o}_t$ is the observation of the speech signal at time t (where an "observation" is described by the features of the speech signal at time t), and k is the k^{th} phonetic symbol associated with $\mathbf{o}_t$. In this framework, both the transitions between states and the phonetic categories within each state are stochastic. As a result, the state occupation sequence is not directly obtainable from the observed speech, but is "hidden"; this hidden stochastic process (or doubly stochastic process) is the reason for the terminology "*hidden* Markov model."

According to this model, the likelihood of a hypothesized utterance is equal to the probabilities of being in each state corresponding to that utterance, multiplied by the probabilities of transitioning between the states in the utterance:

$$P(U) = p(\mathbf{o}_{t1}|q_1) \cdot p(q_2|q_1) \cdot p(\mathbf{o}_{t2}|q_1) \cdot p(q_3|q_2) \cdot \ldots \cdot p(\mathbf{o}_{tN}|q_1) \cdot p(q_N|q_{N-1}) \qquad \text{(A.1)}$$

where U is a hypothesized utterance with states $q_1, q_2, q_3, \ldots q_N$; $p(\mathbf{o}_t|q)$ is the probability of a given speech observation $\mathbf{o}_t$ in state q, which is equal to $b_q(\mathbf{o}_t)$; $p(q_n|q_{n-1})$ is the *a priori* probability of transitioning from the $n-1^{th}$ state to the n^{th} state, which is equal to $a_{n-1,n}$; and N is the number of (possibly non-unique) states in the hypothesized utterance, where each state corresponds to one time frame.

In Figure A.1, each phoneme is associated with one state; in more typical HMM systems, one phoneme is associated with three states, and each state is dependent not only on the current phoneme, but also the previous and next phonemes (a three-state triphone model). Figure A.2 shows one section of a three-state triphone HMM for the word "yes"; due to space considerations, only the states for the phonemes /j/ (/j/ in Worldbet) and /ɛ/ (/E/ in Worldbet) are shown. This figure also shows the probabilities associated with each state at one hypothetical time (or frame), and the probabilities of the utterance being in each state at that time. As the figure indicates, the state with the most likely observation probability (0.96 for **sil-j+E**) does not always correspond to the state with the greatest total likelihood up through the current time (state **j-E+s** in this illustration). The notation "*P-C+N*" in each state is used to represent the current phoneme C in the context of the preceding phoneme P and the next phoneme N.

The process by which an HMM can be used to recognize speech is illustrated in Figure A.3. Short-term spectral-domain features (with a typical window length of 16 msec) are computed from the input speech; often, the delta values of these features are also computed in order to capture some of the dynamics of the speech signal. These features are computed at short, regularly-spaced frames (with 5 to 20 msec per frame), and are usually modified to emphasize the perceptually-relevant aspects of the signal [63, 39]. For classification of a single frame, a context window is taken; this context window includes the features for the current frame and may include features in surrounding frames. In the case where Gaussian mixture models are used to estimate the phonetic likelihoods, the context window usually covers only the one frame of interest, and the delta features are used to implicitly include information about surrounding frames. With an ANN classifier, multiple frames are often used in the context window, in addition to delta features. The frames in the context window are passed to a classifier, which is usually a Gaussian mixture model or a neural network. The classifier estimates the likelihood of each phonetic category at that frame of speech. Classification is done for all frames, and this $F \times C$ matrix of probabilities (where F is the number of frames, and C is the number of categories) is passed to the Viterbi search. The Viterbi search computes the most likely sequence of states, given the phonetic likelihoods at each state (frame), the transition probabilities as determined from the training corpus, and the vocabulary and grammar constraints. The output of the Viterbi search contains not only the most likely word sequence, but also the times at which each state is occupied.

Given this framework, there are four remaining issues: what features are used for classification, how to estimate the observation probabilities, how to estimate the transition probabilities, and how to improve these estimates.

A.2 Features for Classification

The features used by an HMM for classification are called observations. Each observation represents information in the speech signal at one time frame. There are two commonly used feature representations, called PLP and MFCC.

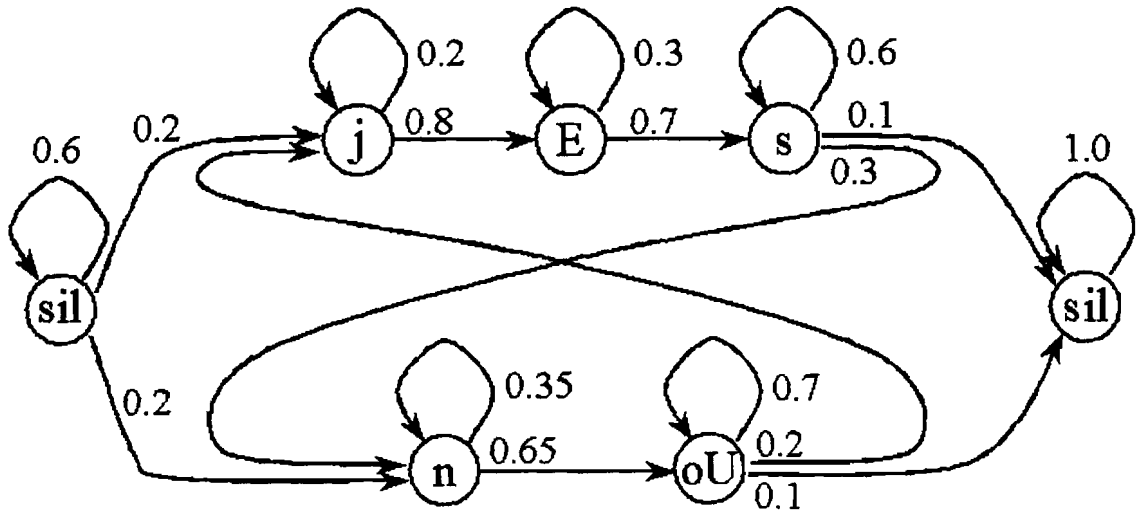

Figure A.1: HMM state sequence for a two-word vocabulary.

Perceptual Linear Prediction (PLP) [63] modifies linear-predictive coding in order to enhance the perceptually-relevant aspects of the signal. Linear-predictive coding (LPC) is a representation of one window of the speech signal; it is usually thought of as a spectral-domain representation. In speech recognition, these LPC coefficients are usually converted to the cepstral domain, in which lower-order coefficients represent low-frequency change in the spectrum (such as spectral tilt), and higher-order coefficients represent high-frequency change in the spectrum (such as the harmonics present in voiced speech). A small number of these cepstral coefficients are used in recognition systems because they are approximately independent of each other, which is advantageous when using a Gaussian mixture model (GMM) to classify the speech sounds, and because a small number of cepstral coefficients can represent the spectral envelope, which contains information about the resonant frequencies of the signal. PLP is related to LPC, except that the model of the speech signal is modified to emphasize the perceptually-relevant aspects. PLP modeling of speech is done in the following stages:

1. windowing of the speech signal around the frame of interest;

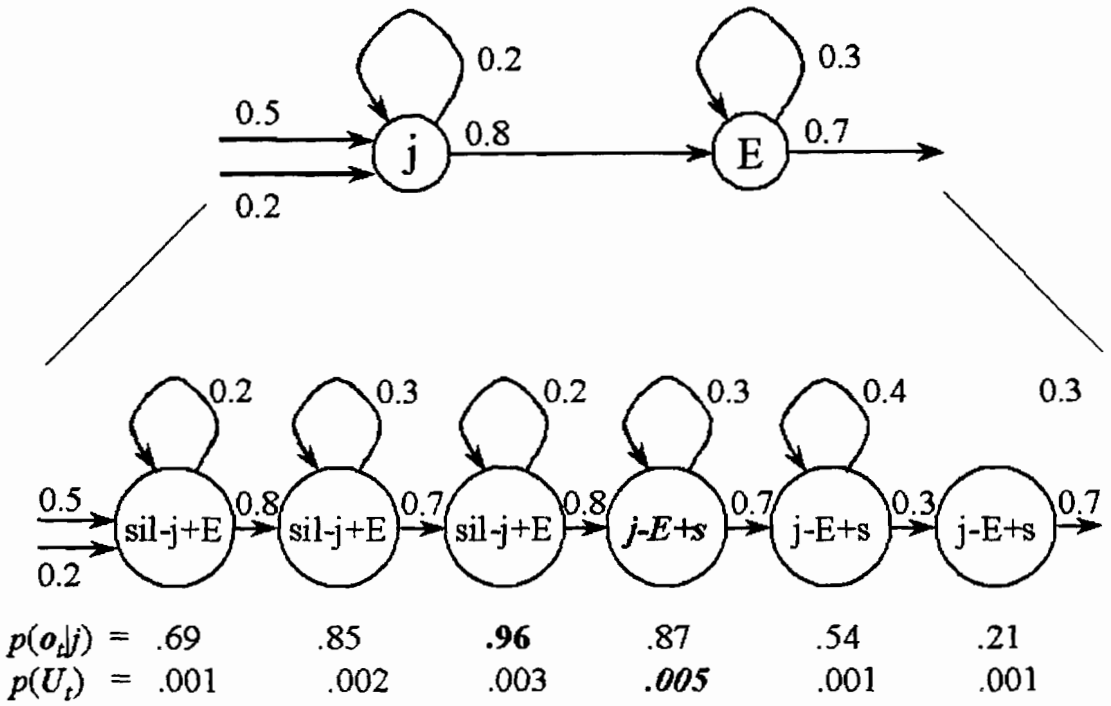

Figure A.2: Expanded HMM state sequence for the first two phonemes (/j/ and /E/) of the word "yes."

2. computation of the power spectrum of the windowed speech, using the FFT;

3. warping the FFT representation along the perceptually-relevant Bark scale;

4. critical-band filtering of the Bark-scale FFT, roughly approximating the properties of human auditory filters;

5. equal-loudness preemphasis of the critical-band filter outputs, compensating for the non-uniform sensitivity of human hearing at different frequencies;

6. amplitude compression of the pre-emphasized filter outputs, approximating the power law of hearing; and

7. LPC modeling of the resulting compressed and pre-emphasized filter outputs.

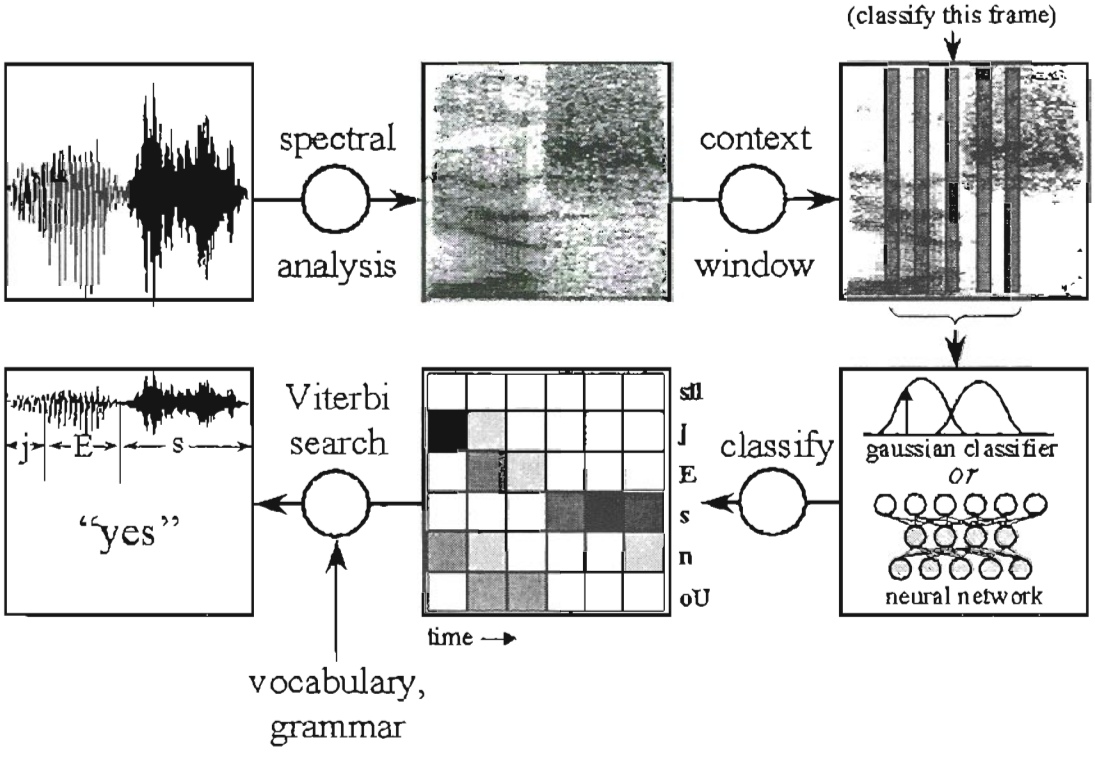

Figure A.3: Graphical overview of the recognition process, illustrating recognition of the word "yes."

8. usually, the LPC coefficients are converted to cepstral coefficients.

Mel-Frequency Cepstral Coefficient (MFCC) features [39] are similar to PLP features in that a perceptual warping of the frequency-domain information is done. However, MFCC does not use LPC modeling of the spectrum, it does preemphasis differently, and it does not do amplitude compression. MFCC modeling is done in the following stages:

1. preemphasis of the windowed speech signal, using a constant factor;

2. windowing of the speech signal around the frame of interest;

3. computation of the power spectrum of the windowed speech, using the FFT;

4. warping the FFT representation along the Mel scale;

5. conversion of the mel-scale representation to cepstral coefficients.

A.3 Estimating the Observation Probabilities

A.3.1 Gaussian Mixture Model Method

In HMM systems that use Gaussian mixture models (GMMs) to estimate the observation probabilities $b_j(\mathbf{o}_t)$, the GMM has the form

$$b_j(\mathbf{o}_t) = \sum_{k=1}^{M} c_{jk}\mathcal{N}(\mathbf{o}_t, \mu_{jk}, \mathbf{U}_{jk}) \tag{A.2}$$

where $b_j(\mathbf{o}_t)$ is the probability of a given speech observation $\mathbf{o}_t$ in state j; c_{jk} is a mixture coefficient for the k^{th} mixture in state j; and $\mathcal{N}$ is a Gaussian probability density function (p.d.f.) for the observation $\mathbf{o}_t$, with mean vector μ_{jk} and covariance matrix $\mathbf{U}_{jk}$ for the k^{th} mixture component in state j.

The initial estimates of $b_j(\mathbf{o}_t)$ can be obtained by one of several methods. According to Rabiner, "experience has shown that good initial estimates are ... essential (when dealing with multiple mixtures) in the continuous-distribution [GMM] case" [120]. For the single-mixture case, one method of obtaining initial estimates of $b_j(\mathbf{o}_t)$ is to associate phonetically-labeled training data with the correct state sequence, and compute the means and covariance matrices of the features for each segment. These means and covariance matrices then specify $b_j(\mathbf{o}_t)$. The phonetically-labeled data can be obtained from manual labels or automatic alignment methods.

A method of obtaining initial values for $b_j(\mathbf{o}_t)$ when using multiple mixtures is based on k-means clustering. First, the training data are segmented into states given the current HMM, using the Viterbi algorithm. From this segmentation, k-means clustering is used to cluster the data for each state j into M mixtures (where M is the number of mixtures per state). The means and covariance matrices of each mixture can be computed, and the mixture weights c_{jm} can be adjusted based on the relative number of data points in each mixture. These new means, covariances, and mixture weights form new estimates of $b_j(\mathbf{o}_t)$.

The transition probabilities a_{ij} can be estimated using the formula

$$\overline{a}_{ij} = \frac{\text{expected number of transitions from state } i \text{ to state } j}{\text{expected number of transitions from state } i} \tag{A.3}$$

with the values in the numerator and denominator obtained from counting occurrences of these events in the training data. These new estimates for $b_j(\mathbf{o}_t)$ and a_{ij} are used to construct a new HMM, and Viterbi segmentation is repeated until the HMM parameters are stable.

Once an initial estimate has been obtained, further refinement is possible, as will be discussed in Section A.4.

A.3.2 Neural-Network Method

The estimates of state occupation can also be determined using a neural-network classifier. Given an accurate alignment of the training data, each observation of a frame of data can be trained to correspond to a given phonetic category. This training will, with sufficient data and nodes in the network, result in output values that are *a posteriori* probabilities of each category or state, given the input data: $p(j|\mathbf{o}_t)$. These probabilities are "discriminant by nature," in that "models are trained to suppress incorrect classification as well as to accurately model each class separately" [11]. In addition, the input features do not have to be independent, allowing the neural network to have a context window of several frames of input, and thereby capturing some dynamic behavior in a way not possible with delta values.

For HMM systems, what $b_j(\mathbf{o}_t)$ should estimate is not, however, $p(j|\mathbf{o}_t)$, but $p(\mathbf{o}_t|j)$. A scaled estimate of $p(\mathbf{o}_t|j)$, called $\hat{p}(\mathbf{o}_t|j)$, can easily be obtained using Bayes' rule, by dividing the neural network outputs by the class prior probabilities:

$$\hat{p}(\mathbf{o}_t|j) = \frac{p(j|\mathbf{o}_t)}{p(j)} \tag{A.4}$$

where $\hat{p}(\mathbf{o}_t|j)$ is related to $p(\mathbf{o}_t|j)$ by

$$\hat{p}(\mathbf{o}_t|j) = \frac{p(\mathbf{o}_t|j)}{p(\mathbf{o}_t)} \tag{A.5}$$

For any observation at time t, $p(\mathbf{o}_t)$ will be the same for all states and will not influence the final result [124]. These scaled likelihoods $\hat{p}(\mathbf{o}_t|j)$ can then be used in the HMM model for $b_j(\mathbf{o}_t)$.

As an alternative to dividing the neural network output by the class priors, the neural network learning algorithm can be modified so that the class prior is approximately "flattened" or factored out during training [145]. The output of such a network is approximately proportional to the *a posteriori* probability divided by the class prior probability, and is called $\dot{p}(j|\mathbf{o}_t)$, where

$$\dot{p}(j|\mathbf{o}_t) \approx \frac{p(j|\mathbf{o}_t)}{p(j)}. \tag{A.6}$$

As a result, $\dot{p}(j|\mathbf{o}_t)$ can then be used in the HMM model as a scaled estimate of $\hat{p}(\mathbf{o}_t|j)$ or $b_j(\mathbf{o}_t)$. Based on several informal experiments, we have found that this modified training provides better word-level recognition results than division by class priors.

A.4 Estimating the Transition Probabilities

The transition probabilities, a_{ij}, can be initially assigned random or (more typically) uniform values. If the k-means method is used to determine initial estimates of $b_j(\mathbf{o}_t)$, then values for a_{ij} are estimated as well. Such methods are "adequate for giving useful reestimates of these parameters in almost all cases" [120]. These initial values can then be iteratively improved, as will be described in Section A.5.

If a_{ij} is used without modification in an HMM system, the duration model has an implicit Geometric distribution, meaning that with increasing time there is notably less likelihood of remaining in a given state [18]. This Geometric model does not fit well with observed duration distributions found in speech data; the data are better fit by a Gamma distribution. To resolve this issue, the structure of the HMM can be modified to better approximate the true duration distributions [121]. Another simple but effective method is to replace the standard a_{ij} probabilities by penalties during the Viterbi search. The value of the penalty depends on how long a state is occupied. In the Burshtein model [18], a true Gamma distribution is obtained; in the baseline CSLU recognizers [68], a penalty is applied if the state duration is too short or too long, but no penalty is applied for "typical"

durations.

A.5 Updating the Probability Estimates

Given initial estimates of a_{ij} and $b_j(\mathbf{o}_t)$ from the training data, these estimates can be improved using the "forward-backward" method. The forward-backward method is an iterative procedure that takes as input an initial HMM and observation sequence. Given these parameters, we can compute the probability of a partial observation sequence from time t_1 to time t_t (corresponding to $\mathbf{o}_1\mathbf{o}_2\mathbf{o}_3\ldots\mathbf{o}_t$) and ending in state i at time t_t, using a recursive procedure. This probability is called the forward probability, and is denoted $\alpha_t(i)$. In a similar way, the partial observation sequence going backward in time from $t = T$ to $t = t+1$ ($\mathbf{o}_T\mathbf{o}_{T-1}\ldots\mathbf{o}_{t+1}$) and ending in state j can be computed (where T is the final time); this probability, called the backward probability, is denoted $\beta_{t+1}(j)$. Given $\alpha_t(i)$, $\beta_{t+1}(j)$, a_{ij}, and $b_j(\mathbf{o}_{t+1})$, we can compute the probability of being in state i at time t and in state j at time $t+1$, which is called $\xi_t(i,j)$. From $\xi_t(i,j)$, we can compute the probability of being in state i at time t given the entire observation sequence, which is called $\gamma_t(i)$. If we define re-estimation formulae for a_{ij} and $b_j(k)$ (in this case, defining $b_j(k)$ for discrete categories instead of continuous probability distribution functions, although the same ideas apply) as

$$\bar{a}_{ij} = \frac{\text{expected number of transitions from state } i \text{ to state } j}{\text{expected number of transitions from state } i} \tag{A.7}$$

$$= \frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)} \tag{A.8}$$

$$\bar{b}_j(k) = \frac{\text{expected number of times in state } j, \text{ observing symbol } \mathbf{v}_k}{\text{expected number of times in state } j} \tag{A.9}$$

$$= \frac{\sum_{\substack{t=1 \\ \text{s.t.}\mathbf{o}_t=\mathbf{v}_k}}^{T}\gamma_t(j)}{\sum_{t=1}^{T}\gamma_t(i)} \tag{A.10}$$

then we can use $\xi_t(i,j)$ and $\gamma_t(i)$ to re-estimate a_{ij} and $b_j(k)$. This re-estimation can be applied iteratively to generate new estimates of a_{ij} and $b_j(k)$ until a local maximum-likelihood estimate is obtained.

The forward-backward method can be used to achieve improvement in recognition performance with either the Gaussian-mixture-model or neural-network approaches [122, 153].

Appendix B

Glossary of Speech Terminology

A* search: a heuristic search that is able to return the N-best search paths through a network of states.

acoustic-phonetic information: distinct events in the speech signal (acoustics) that correspond to specific phonetic characteristics

acute: a high value for the average of the first two formants, or a high second formant.

affricate: a phoneme composed of a stop followed immediately by a fricative, such as /ʧ/ or /ʤ/.

approximant: sounds produced in which the articulators are moderately close (without producing frication), such as /j/, /l/, /ɻ/, and /w/.

articulators: the physical parts of the speech-production mechanism, such as the lips or tongue.

autocorrelation: a measurement of how closely a segment of a signal matches the signal later in time; this measurement can be used to determine periodicity in a signal.

Bark scale: a perceptually-based warping of the frequency scale emphasizing the lower frequency regions. Similar to the Mel scale.

bigram: two adjacent words (or phonemes, in the case of phoneme-level recognition)

burst: see *plosive*.

channel: according to Webster's dictionary, a path along which information passes in the form of an electrical signal. Typical channels in speech recognition include head-mounted or stand-alone microphones, land-line telephones, or cellular telephones.

closure: silence or the silent region preceding a burst.

coarticulation: the effect that one phoneme has on its neighboring phonemes; this effect is manifested as a smooth change in formant frequencies from one phoneme to the next.

compact: in some references, *compact* means a sound with frequencies concentrated in one region of the spectrum, computed as the difference between the first and second formants; in other references, compact indicates a high first formant.

critical band: a pass-band filter in the auditory system.

CVC: a consonant-vowel-consonant sequence.

deltas, delta values: a measure of change in a parameter; note that delta values are usually not computed by simply taking the difference between successive values, but using a function that estimates the rate of change with a context window of several values.

dendrogram: a hierarchical representation of segments in a speech signal, with longer segments at the top of the hierarchy, and higher-level segments divided into subsegments at lower levels.

diffuse: in some references, diffuse means a sound with frequencies spread across the spectrum, computed as the difference between the first and second formants; in other references, diffuse indicates a low first formant.

diphone: the region from the middle of one phoneme to the middle of the next phoneme, or alternatively, the transition region between two phonemes.

distinctive (phonetic) features: according to Ladefoged, a distinctive feature is "a phonetic property that can be used to classify sounds." This broad definition allows a large number of possible distinctive features. Common features include voicing, manner, place, and height.

dynamic time warping (DTW): a dynamic-programming technique to align two similar signals in time

forced alignment: automatic alignment of the phonemes in an utterance by constraining the search in a phonetic recognizer to the known sequence of phonemes.

formant: an energy resonance at a particular frequency that is a direct result of a voiced sound source passing through the vocal tract. There are a number of formants in

voiced speech, which are typically referred to by number, with the first formant being the one lowest in frequency. The first three formant frequencies of a neutral vowel are typically located at 500, 1500, and 2500 Hz.

frication: sound produced by forcing air through a narrow constriction at the lips or along the vocal tract.

fricative: a sound that is produced by forcing air through a narrow constriction at the lips or along the vocal tract, such as /f/, /s/, and /ʃ/.

fundamental frequency (F0): the rate at which the vocal folds vibrate during voiced speech.

fundamental period: the inverse of the fundamental frequency, which is equivalent to the time from one pitch-related pulse to the next

glide: sounds such as /ɻ/ or /j/.

glottalization: aperiodic or extremely slow vibration of the vocal folds, which sometimes occurs at word boundaries.

grapheme: a unit of a writing system, such as a letter or character.

grave: a low value for the average of the first two formants, or a low second formant.

harmonic: according to Webster's dictionary, "a component frequency of a complex wave ... that is an integral multiple of the fundamental frequency"

height: a distinctive feature that classifies phonemes according to the vertical position of the tongue in the mouth. The Height feature is related to the location of the first formant.

inverse filter: a filter that removes the effects of its counterpart filter; for example, an inverse filter for a 200-Hz low-pass filter would be a 200-Hz high-pass filter.

Kohonen network: a neural network that is capable of unsupervised learning. Also called a "self-organizing map."

larynx: the upper part of the trachea containing the vocal cords.

lateral: articulation in which the air flows around the sides of the tongue, as in the /l/ sound.

linear-predictive coding (LPC): a representation of a signal in terms of filter coefficients.

liquid: sounds such as /l/ or /w/.

manner: a distinctive feature that classifies phonemes according to their sound source (as in vowels or fricatives) or special positions of the articulators (as in approximants).

Mel scale: a perceptually-based warping of the frequency scale, emphasizing the lower frequency regions. Similar to the Bark scale.

multivariate gaussian: a Gaussian (or Normal) probability distribution in multiple dimensions.

nasal: a sound produced with airflow through the nose, such as /m/, /n/, or /ŋ/.

observation: the output of a state in an HMM, described by the features of the speech signal at a given time.

obstruent: a sound that is a plosive, affricate, or fricative.

offglide: the final portion of a diphthong, such as the /i/ in /ei/

phoneme: a unit of a spoken language that is perceived to be a single sound in that language

phonotactics: according to Webster's dictionary, "the area of phonology concerned with the analysis and description of the permitted sound sequences of a language"

pitch period: the time from one pitch-related pulse to the next, which is equivalent to the inverse of the fundamental frequency

place: a distinctive feature that classifies phonemes according to the horizontal position of the tongue in the mouth (as in front, middle, or back). The Place feature is related to the degree of separation of the first and second formants.

plosive: a sound produced by a buildup of air pressure behind a constriction in the mouth, followed by sudden release of the constriction, as in /p^h/, /t^h/, and /k^h/.

power spectrum: the energy (power) of a signal at each frequency for one time frame.

retroflex: a sound produced with the tip of the tongue and back part of the alveolar ridge, such as /ɻ/.

sonorant: sonorant a sound that is a nasal, liquid, or glide.

spectral envelope: the overall shape of the spectrum, as indicated by the peaks of each harmonic. The spectral envelope contains information about a voice's formants, whereas the harmonics contain information about the fundamental frequency.

spectrogram: a display of the characteristics of a signal in terms of time, frequency, and energy.

"steady-state" region: the region of a phoneme that is not as much influenced by coarticulation. In fluent speech, most phonemes are influenced throughout by coarticulation, in which case the steady-state region is the region that best characterizes the phoneme (as opposed to the phonetic transition regions).

stop: according to Ladefoged, a *stop* is "complete closure of two articulators", such as happens with plosives.

VCV: a vowel-consonant-vowel sequence.

vector quantization: representation of a vector that may take on continuous values using a finite-sized codebook of vectors.

velum: the soft palate (the soft structure at the roof of the mouth).

Viterbi search: a dynamic-programming search that finds the single most-likely path through a sequence of states, based on each state's occupation probabilities and transition probabilities.

voice bar: a low-frequency, periodic resonance that sometimes occurs prior to voiced plosives. This resonance is the result of the vocal folds vibrating before the constriction in the vocal tract is removed.

voiced: a speech signal that has periodic vibration of the vocal folds (voicing).

voice-onset time: the time from the impulse in a plosive until the beginning of voicing in a plosive-vowel phoneme pair.

voicing: a measure of periodicity in the waveform that occurs when the vocal folds vibrate.

Appendix C

Worldbet and IPA Phonetic Symbols

Worldbet and IPA Phonetic Symbols for American English:

Worldbet, IPA		Example
i:	iː	beet
I	ɪ	bit
E	ɛ	bet
@	æ	bat
I_x	ɨ	roses
u_x	ʉ	suit
&	ə	above
&_0	ə̥	to go
u	u	boot
U	ʊ	book
^	ʌ	above
>	ɔ	caught
A	ɑ	father
3r	ɝ	bird
&r	ɚ	butter
ei	ei	bay
aI	aɪ	bye
>i	ɔi	boy
iU	iʊ	few
aU	aʊ	about
oU	oʊ	boat

Worldbet, IPA		Example
ph	p^h	pan
th	t^h	tan
kh	k^h	can
b	b	ban
d	d	dan
g	g	gander
m	m	me
n	n	knee
N	ŋ	sing
th_(	ɾˀ	writer
d_(	ɾᵈ	rider
n_(	ɾⁿ	winner
f	f	fine
T	θ	thigh
s	s	sign
S	ʃ	shine
h	h	hope
h_v	ɦ	she had
v	v	vine
D	ð	this
z	z	resign

Worldbet, IPA		Example
Z	ʒ	azure
tS	ʧ	church
dZ	ʤ	judge
l	l	lent
9r	ɻ	rent
j	j	yes
w	w	went
m=	m̩	bottom
n=	n̩	button
N=	ŋ̩	easing
pc	—	_pan
tc	—	_tan
kc	—	_can
bc	—	_ban
dc	—	_dan
gc	—	_gander
tSc	—	_church
dZc	—	_judge
_?	ʔ̞	*(glottalized)*
.br	—	*(breath)*
.pau	—	*(closure)*

Biographical Note

John-Paul Hosom was born on June 9, 1965 in Manhassett, New York. He attended Rensselaer Polytechnic Institute from 1983 through 1984, and then transferred to the University of Massachusetts at Amherst beginning in 1985. He received his bachelor of science degree in Computer and Information Science from the University of Massachusetts in 1987, graduating *cum laude*. Having studied the Japanese language for two years while at the university, he decided it would be worthwhile to go to Japan before beginning his professional career. Once in Japan, he taught conversational English for two years. After that, he was fortunate enough to be employed by Sumitomo Electric Industries, Ltd. in Osaka, in their research and development department. He worked for four years at Sumitomo on formant-based speech synthesis before returning to the United States. Paul entered the Oregon Graduate Institute of Science and Technology in the Fall of 1994, and has been busy since that time working on speech recognition, automatic phonetic alignment, tutorials, and visual display tools. In 1997, he won first place in the "oral presentations" category of the Student Research Symposium, and second place in the "written papers" category. He also served on the Student Council for two years, from 1996 to 1998. He was invited to Padova, Italy in September 1999 for three weeks, where he worked at the National Research Council's Institute of Phonetics with Piero Cosi on Italian speech recognition. Paul is an author on 11 conference proceedings, 3 journal articles, and one U.S. Patent (no. 5,577,160).