Acoustic Landmark Detection and Segmentation using the McAulay-Quatieri Sinusoidal Model

by

Tara N. Sainath

B.Sc., Massachusetts Institute of Technology (June 2004)

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

Masters of Engineering in Computer Science and Electrical Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

[Se... ...]
August 2005

Author .....
Department of Electrical Engineering and Computer Science
August 16, 2005

Certified by .....
Timothy J. Hazen
Research Scientist, CSAIL
Thesis Supervisor

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

Acoustic Landmark Detection and Segmentation using the McAulay-Quatieri Sinusoidal Model

by

Tara N. Sainath

Submitted to the Department of Electrical Engineering and Computer Science
on August 16, 2005, in partial fulfillment of the
requirements for the degree of
Masters of Engineering in Computer Science and Electrical Engineering

Abstract

The current method for phonetic landmark detection in the Spoken Language Systems Group at MIT is performed by SUMMIT, a segment-based speech recognition system. Under noisy conditions the system's segmentation algorithm has difficulty distinguishing between noise and speech components and often produces a poor alignment of sounds. Noise robustness in SUMMIT can be improved using a full segmentation method, which allows landmarks at regularly spaced intervals. While this approach is computationally more expensive than the original segmentation method, it is more robust under noisy environments. In this thesis, we explore a landmark detection and segmentation algorithm using the McAulay-Quatieri Sinusoidal Model, in hopes of improving the performance of the recognizer in noisy conditions.

We first discuss the sinusoidal model representation, in which rapid changes in spectral components are tracked using the concept of "birth" and "death" of underlying sinewaves. Next, we describe our method of landmark detection with respect to the behavior of sinewave tracks generated from this model. These landmarks are interconnected together to form a graph of hypothetical segments. Finally, we experiment with different segmentation algorithms to reduce the size of the segment graph.

We compare the performance of our approach with the full and original segmentation methods under different noise environments. The word error rate of original segmentation model degrades rapidly in the presence of noise, while the sinusoidal and full segmentation models degrade more gracefully. However, the full segmentation method has the largest computation time compared to original and sinusoidal methods. We find that our algorithm provides the best tradeoff between word accuracy and computation time of the three methods. Furthermore, we find that our model is robust when speech is contaminated by white noise, speech babble noise and destroyer operations room noise.

Thesis Supervisor: Timothy J. Hazen
Title: Research Scientist, CSAIL

Acknowledgments

I would first like to express my deepest gratitude to my thesis advisor T.J. Hazen. His patience and mentorship over the past year has helped guide me through this thesis. Furthermore, his explanations and insightful suggestions have helped shape me as a researcher.

I would also like to acknowledge Victor Zue and the research staff for welcoming me into the Spoken Language Systems Group and creating a supportive and simulating research environment. Also thank you to Lee Hetherington for his help with the recognizer.

I would like to thank my friends and family for their unwavering love and support; To all the friends I have made over the past 5 years, thank you for making my time at MIT truly rewarding and enjoyable. Thank you to my sister for her patience and her companionship. Thank you to my uncle, aunt and cousins for their constant encouragement and my adorable nieces and nephew for always making me laugh.

Finally, I would like to my parents for being wonderful role models and for always being a source of inspiration in my life.

Contents

1 Introduction **14**

1.1 Problem Statement and Motivation 14

1.2 Noise Robust Speech Recognition 16

1.2.1 Previous Work 16

1.2.2 Proposed Noise Robust Technique 18

1.3 Thesis Goals 18

1.4 Overview 19

2 System Components **20**

2.1 Speech Recognition Corpora 20

2.1.1 AV-TIMIT 20

2.1.2 Noisex-92 21

2.2 SUMMIT Speech Recognition System 22

2.2.1 Mathematical Formulation 22

2.2.2 Acoustic Model 23

2.2.3 Pronounciation/Lexical Model 24

2.2.4 Language Model 24

2.2.5 Recognition Phase 25

3 Sinusoidal Modeling of Speech **27**

3.1 Acoustic Theory of Speech Production 27

3.2 The McAulay-Quatieri Algorithm 28

3.2.1 Analysis 28

3.2.2 Peak-to-Peak Matching 29
3.2.3 Synthesis 30
3.2.4 Improvements to Original MQ Algorithm 30
3.3 Other Sinusoidal Modeling Techniques 32
3.3.1 Phase Vocoder 32
3.3.2 Spectral Modeling Synthesis 32
3.3.3 Harmonic Plus Noise Model 33

4 Landmark Detection **35**
4.1 Sinusoidal Model 35
4.2 Endpoint Location Method 37
4.2.1 Short-Time Energy 38
4.2.2 Harmonicity 40
4.3 Detecting Landmarks from Sinusoidal Components 43
4.3.1 Identifying Harmonically Related Sinusoids 44
4.3.2 Landmarks from Harmonic Sinusoids 45
4.3.3 Landmarks in Unvoiced Regions 46

5 Segmentation **47**
5.1 Background 47
5.2 Full Segmentation 48
5.3 Original Segmentation 48
5.4 Sinusoidal Model Segmentation 50
5.4.1 Segmentation at Voiced/Unvoiced Boundaries 52
5.4.2 Segment Connectivity Methods 53
5.4.3 Segmentation Using MFCC Distance Information 54

6 Experimental Results **60**
6.1 Experimental Setup 60
6.2 Sinusoidal Model Landmarks 61
6.2.1 Landmark Detection Parameter Settings 61

6.2.2 Phonetic Detection Probability 64
6.2.3 Landmark Types Not Used 66
6.3 Sinusoidal Model Segmentation Methods 67
6.3.1 Voicing Decision Methods 68
6.3.2 Connectivity Methods 70
6.3.3 MFCC Distance 70
6.4 Comparison of Models 71
6.4.1 Word-Error-Rate 71
6.4.2 Recognition Computation Time 76
6.4.3 Word Error Rate vs. Computation Time 78
6.4.4 Landmark and Segment Comparisons 82

7 Conclusions **86**
7.1 Summary 86
7.1.1 Algorithm Design 86
7.1.2 Performance of Sinusoidal Model 87
7.2 Future Work 88

A AV-TIMIT Phonemes **90**

B Word Error Rate Tables **91**
B.1 Word Error Rate for White Noise 92
B.2 Word Error Rate for Babble Noise 93
B.3 Word Error Rate for Destroyerops Noise 94

List of Figures

1-1 Block diagram of a segment-based speech recognition system 15

3-1 Block Diagram of Analysis Component [19] 29

3-2 Birth and Death of Sinusoidal Tracks [4] 30

3-3 Block Diagram of Synthesis Component [19] 31

4-1 Block Diagram of Landmark Detector 35

4-2 Typical sinusoidal tracks for voiced and unvoiced speech overlaid on a speech spectogram 37

4-3 Sinusoidal Track Representation and Corresponding Signal Energy of Speech Utterance 39

4-4 Sinusoidal Track Representation and Corresponding Harmonicity of Speech Utterance 41

4-5 Sinusoidal Tracks and Harmonicity of Speech with a SNR of -5db. The speech is contaminated by white noise. 42

5-1 Segment network for Full Segmentation technique. Each landmark l_i is fully connected to every other landmark l_j in the graph via segments s_{ij}. 48

5-2 Graphical display of the segment network for the full segmentation approach from the SUMMIT recognizer 49

5-3 Segment network for Original Segmentation technique. Major landmarks are indicated in bold. Each minor landmark l_i between major landmarks is fully connected to every other landmark l_j in the graph via segments s_{ij}. In addition, each major landmark is connected to two major landmarks forward. 50

5-4 Graphical display of the segment network for the original segmentation approach from the SUMMIT recognizer 51

5-5 Segment network for Sinusoidal Model Two-Connection technique. Major landmarks are indicated in bold. Each minor landmark l_i is connected to every other landmark l_j which falls up to two major landmarks away via segments s_{ij}. In addition, each major landmark is connected to the next three major landmarks. 54

5-6 Segment network for Partial-Connection technique. Hard major landmarks are indicated in bold while soft major landmarks are embossed. Minor landmarks l_i are connected across a soft major landmarks to other landmarks l_j which falls up to two major landmarks away via segments s_{ij}. However, minor landmarks cannot be connected across hard major landmarks. In addition, each major landmark is connected to the next three major landmarks. 55

5-7 Graphical display of the segment network for the partial-connection method from the SUMMIT recognizer 55

5-8 MFCC feature vector distance matrix of a speech signal. 57

5-9 The top figure shows a major landmark overlaid on top of the speech signal. Notice the large change in MFCC distance on either side of the major landmark, as illustrated in by the bottom figure. 58

5-10 The top figure shows a major landmark overlaid on top of the speech signal. The bottom figure illustrates a small change in MFCC distance on either side of the major landmark. 59

6-1 Example Receiver Operating Characteristic (ROC) Curve plotting the Probability of Detection, P_d versus the Probability of False Alarm, P_{fa} 62

6-2 Generated Receiver Operating Characteristic (ROC) Curve for Landmark Detection Parameters 63

6-3 AV-TIMIT phoneme detection probability under clean speech 65

6-4 Recognition computation time for different segmentation approaches under varied SNR of white noise 68

6-5 Word Error Rate for Original Segmentation, Full Segmentation and Sinusoidal Models for varied SNRs of White Noise 73

6-6 Word Error Rate for Original Segmentation, Full Segmentation and Sinusoidal Models for varied SNRs of Babble Noise 74

6-7 Word Error Rate for Original Segmentation, Full Segmentation and Sinusoidal Models for varied SNRs of Destroyerops Noise 75

6-8 Recognition computation time for original, full and sinusoidal methods under varied SNR of white noise 77

6-9 Word Error Rate vs. Computation Time for Original, Full and Sinusoidal Methods. Results are computed for varied *vprunenodes* parameter under 5dB of white noise. 79

6-10 Word Error Rate vs. Computation Time for Original, Full and Sinusoidal Methods. Results are computed for varied *vprunenodes* parameter under 5dB of babble noise. 80

6-11 Word Error Rate vs. Computation Time for Original, Full and Sinusoidal Methods. Results are computed for varied *vprunenodes* parameter under 5dB of destroyerops noise. 81

6-12 Average time difference between a hypothesized and AV-TIMIT landmark for original, full and sinusoidal methods under white noise . . . 83

6-13 Average number of segments across an AV-TIMIT phoneme for original, full and sinusoidal methods under white noise 84

A-1 61 AV-TIMIT phones and corresponding International Phonetic Alphabet (IPA) Symbols, along with example words using the phonemes 90

List of Tables

6.1 Word Error Rates for Different Sinusoidal Model Segmentation Approaches for varied SNRs of White Noise. Bold represents best segmentation method for each noise condition 69

6.2 Comparison matrix showing results of McNemar's Test for Sinusoidal Model, Original Segmentation and Full Segmentation methods under varied SNRs of white noise. Methods which are statistically similar are indicated with an $\approx$ symbol and the corresponding significance level. If two methods are statistically different, the model with the better performance is indicated. Also, the model with the lowest error rate for each noise condition is indicated in bold. 72

B.1 Word Error Rates for Original, Full and Sinusoidal Approaches for varied SNRs of White Noise. Bold represents best method for each noise condition. 92

B.2 Word Error Rates for Original, Full and Sinusoidal Approaches for varied SNRs of Babble Noise. Bold represents best method for each noise condition. 93

B.3 Word Error Rates for Original, Full and Sinusoidal Approaches for varied SNRs of Destroyerops Noise. Bold represents the best method for each noise condition. 94

Chapter 1

Introduction

Speech is subject to contamination from many different noise sources. Humans are able to differentiate this speech from noise and comprehend the spoken words. While speech recognition systems are able to successfully process human speech, their performance degrades rapidly in the presence of background noise. The purpose of this research is to explore a technique to improve the noise robustness of a segment-based speech recognition system.

1.1 Problem Statement and Motivation

Most speech recognition systems represent a speech utterance by a temporal sequence of frame-based feature vectors. To date, Hidden Markov Models (HMMs) have been the most dominant frame-based acoustic modeling technique for automatic speech recognition. Although HMMs have proven to be very successful in many tasks, alternative models have also been developed to address the limitations of HMMs [22].

One type of alternative model that has been developed is a segment-based speech recognizer. The SUMMIT speech recognizer developed by the Spoken Language Systems group at MIT uses a segment-based framework for acoustic modeling [10]. Figure 1-1 shows a block diagram of the SUMMIT recognition system. SUMMIT computes a temporal sequence of frame-based feature vectors from the speech signal, but then hypothesizes acoustic landmarks at regions of large change within these feature vectors

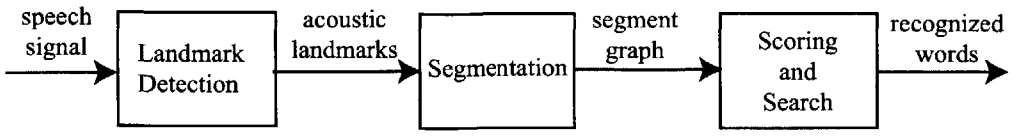

Figure 1-1: Block diagram of a segment-based speech recognition system

[9]. These landmarks represent possible transitions between phones.

These landmarks are then connected together to form a graph of possible segmentations of the utterance. To minimize the number of interconnections among landmarks, an explicit set of segmentation rules is incorporated into SUMMIT to reduce the size of the segment graph. This graph is passed to scoring and search components which use frame and segment-based measurements to score phonetic hypotheses and find the optimal path of phonetic elements through the segment graph. We will refer to this baseline segmentation algorithm used by SUMMIT as the *original segmentation* method in this thesis.

While SUMMIT is able to process human speech in noise-free environments, the system's segmentation algorithm performs poorly in the presence of strong background noises and non-speech sounds. Specifically, the system has a difficult time distinguishing between the noise and speech components and often produces a poor alignment of sounds.

Noise robustness in SUMMIT can be improved using a *full segmentation* method. This technique places landmarks at equally spaced intervals and outputs a segment graph which fully interconnects all landmarks. While this approach is computationally more expensive than the original segmentation method, it is more robust under noisy environments.

In this thesis, we investigate a new segmentation algorithm to improve the performance of the recognizer under contaminated speech. In addition, we explore alternative segmentation approaches to reduce computation time but still allow for noise robustness.

1.2 Noise Robust Speech Recognition

In recent years, improvements in speech recognition systems have resulted in high performance of specific tasks under clean conditions. For example, digit recognition has resulted in a less than a 0.5% word error rate. In addition, an error rate of less than 1% has been achieved in an isolated word recognition task [12].

However, the performance of these systems rapidly degrades in noisy environments. For example, the performance of a recognizer in a clean speech environment can drop by over 30% in accuracy when the same speech is corrupted over long-distance telephone lines [20].

Various phenoma occur under noisy conditions which can explain the degradation of these systems [16]. Additive noise alters the speech signal and hence the feature vectors used by speech recognizers to represent this signal. For example, white noise has been shown to reduce the variance of cepstral coefficients [31].

1.2.1 Previous Work

To date, it has not been possible to develop a universally successful and robust speech recognition system for all environmental conditions. Systems which perform well in one scenario can seriously degrade in performance under a different environmental stress. Therefore, there have been numerous techniques studied to improve the robustness of speech systems under noisy conditions [12]. These techniques can be divided into three main categories based on their objectives:

- Noise Resistant Features
- Speech and Feature Enhancement
- Noise Adaptation

Noise Resistant Features

Methods in this category attempt to use features which are less sensitive to noise and distortion. These methods focus on identifying better speech recognition features or

estimating robust features in the presence of noise. While these techniques do not make assumptions nor estimations about noise characteristics, this is sometimes a disadvantage since it is impossible to fully utilize features which are specific to a noise type.

Speech Enhancement

Speech enhancement can be used as a preprocessing step for recognition. These methods attempt to suppress the impact of noise on speech by extracting out clean speech or feature vectors from a contaminated signal. Some approaches include parameter mapping, spectral subtraction, noise masking, comb filtering, Bayesian estimation and parametric spectral modeling. While these techniques are capable of improving recognition performance, oftentimes while the estimated clean speech appears more intelligible to humans it does not necessarily show improvement in the recognizer. Furthermore, it is sometimes difficult to develop a speech enhancement technique capable of suppressing a multitude of noise types.

Noise Adaptation

Instead of deriving an estimate of clean speech, noise adaptation techniques attempt to adapt recognition models to noisy environments. This includes changes to the recognizer formulation, such as changing model parameters of the recognizer, to accommodate noisy speech. Parallel Model Combination [7] is one such method for compensating model parameters under noisy conditions in a computationally efficient manner. In addition, some techniques also explore designing noise models within the recognizer itself. While this technique performs well at high SNRs, at low SNRs compensated model signal parameters often show large variances, resulting in a rapid degradation of performance.

1.2.2 Proposed Noise Robust Technique

The previous noise robustness techniques all incorporate an extra stage into the recognition process in an attempt to clean up contaminated speech and to improve the scoring stage of the recognition process. In this thesis, we attack an orthogonal problem of improving the segmentation phase of the SUMMIT recognition system under noisy conditions. We do not address the problems posed by noise in the feature extraction and acoustic modeling components of the system.

The McAulay-Quatieri Sinusoidal Modeling Algorithm, developed by Tom Quatieri and Robert McAulay [19], models a periodic signal as a collection of sinusoidal components. Representing a speech signal via this sinusoidal model can help to separate out the harmonic speech components from residual aperiodic noise. Rapid changes in spectral components are tracked using the concept of "birth" and "death" of the underlying periodic sine waves.

We will detect landmarks by looking at the births and deaths of these sinusoids. These landmarks are hypothesized from the sinusoidal behavior of the contaminated speech itself, as opposed to detecting features after applying a speech enhancement technique or using alternative noise resistant features. Landmarks are connected to form a segment graph, an appropriate segmentation algorithm is applied to reduce the size of the search space, and the optimal sequence of phonemes is found based on segment-based measurements derived from the contaminated speech itself.

1.3 Thesis Goals

The overall goal in this thesis is to develop an appropriate landmark detection and segmentation algorithm which provides a good tradeoff between word error rate and computation time under different noise environments. More specifically, one goal of this thesis is to develop a robust landmark detection method that will lead to an improvement in word recognition accuracy over the original segmentation method. Another goal is to develop an appropriate segmentation algorithm to provide faster computation time over the full segmentation approach. In our approach, we detect

landmarks and develop a segmentation algorithm based on the behavior of sinusoidal tracks derived from noisy speech. We hope that our method will be robust to many different noise conditions, a limitation of many previous noise robust techniques.

1.4 Overview

The remainder of this thesis is organized in the following manner. Chapter 2 describes the speech corpora and segment-based system used used for recognition experiments in this thesis. Sinusoidal modeling techniques, including the McAulay-Quatieri Sinusoidal Modeling Algorithm, will be discussed in Chapter 3. Hypothetical landmarks are detected via a landmark detection method, which is described in Chapter 4. Chapter 5 discusses numerous segmentation approaches for the full segmentation, original segmentation and sinusoidal model methods. Chapter 6 compares the word error rate and computation time of the three methods. Finally, Chapter 7 concludes the work in this thesis and provides a few remarks about future work.

Chapter 2

System Components

This chapter describes the speech corpora and segment-based system used used for recognition experiments in this thesis.

2.1 Speech Recognition Corpora

Our recognition experiments use the AV-TIMIT corpus. This corpus contains phonetically rich and varied audio-visual recordings of read speech. Noisy speech is then simulated by adding noises from the Noisex-92 corpus to the utterances from AV-TIMIT. The following sections describe the corpora in more detail.

2.1.1 AV-TIMIT

The Audio-Visual TIMIT (AV-TIMIT) corpus [14] is a collection of speech recordings developed at the Massachusetts Institute of Technology for research in audio-visual speech recognition. The speech data in AV-TIMIT was recorded with a far-field array microphone and video camcorder in a quiet, controlled office setting. Although the full corpus contains both audio and visual data, the work in this thesis uses only the audio data. One of the main design goals for the corpus was to create a phonetically balanced collection of speech utterances. The TIMIT-SX collection was used to provide a wide range of phonetic contexts of the English language [32]. In

total, 23 different rounds of utterances were created. In each round, a set of 20 sentences is read by a speaker. The first sentence in each round is the same for all speakers to adapt them to recording process, while the other 19 sentences for each round differ. The final corpus contains 223 speakers, including 117 males and 106 females. The total database duration is approximately 4 hours. The sentences from the corpus are divided into three sets:

- The *train* set consists of 3793 sentences. This is used to train various models used by the recognizer.
- The *dev* set contains 399 sentences. We used 285 of these utterances to design and develop the sinusoidal model.
- The *test* set includes 405 sentences. We used 299 of these utterances use to test our developed model.

To create unbiased experimental conditions, the sentences in the *train*, *dev* and *test* sets do not overlap.

2.1.2 Noisex-92

The Noisex-92 speech-in-noise database [30] was created by the Speech Research Unit at the Defense Research Agency to study the effect of additive noise on speech recognition systems. The database contains the following noises:

- White Noise, Pink Noise, High Frequency Radio Channel Noise
- Speech Babble
- Factory Noise
- Military Noises: fighter jets (Buccaneer, F16), destroyer noises (engine room, operations room), tank noise (Leopard, M109), machine gun
- Volvo 340 Car Noise

In this thesis we look at three specific types of noises, white-noise, speech babble and destroyer operations room noise. The white noise was acquired by sampling a high-quality analog noise generator. The speech babble was obtained by recording samples of 100 people speaking in a canteen. Finally, the destroyer operations room noise was obtained by recording noise samples in an operation room onto a digital audio tape.

We simulate noisy speech by adding noise from the Noisex-92 set to clean AV-TIMIT speech at signal-to-noise ratios in the range of -10db and 20db.

2.2 SUMMIT Speech Recognition System

SUMMIT is a segment-based speech recognition system developed at the Spoken Language Systems Group at MIT's Computer Science and Artificial Intelligence Laboratory [10]. In this section we will briefly discuss the different components of the SUMMIT recognition system.

2.2.1 Mathematical Formulation

Given a set of acoustic observations $A = \{a_1, a_2, a_3, \ldots, a_n\}$ associated with a speech waveform, the goal of a speech recognition system is to find the corresponding sequence of words $\hat{W} = \{w_1 w_2 ... w_n\}$ which has the maximum *a posteriori* probability $P(W|A)$. This goal is expressed more formally by Equation 2.1:

$$\hat{W} = \arg\max_{W} P(W|A) \tag{2.1}$$

In a segment-based recognition system, multiple segmentations, S, are associated with the acoustic observations. For each segmentation, there is an associated sequence of sub-word units U. Sequences of these sub-word units form a corresponding sequence of words, W. Taking into account the segmentation and sub-word units, Equation

2.1 can be rewritten as

$$\hat{W} = \arg\max_{W} \sum_{S} \sum_{U} P(W, U, S|A) \tag{2.2}$$

To simplify computation, SUMMIT uses dynamic programming (e.g., Viterbi) or graph-searches (e.g., A^*) to find a single optimal segmentation $\hat{S}$, along with an optimal unit sequence $\hat{U}$ and words $\hat{W}$. Equation 2.2 then simplifies to:

$$\hat{W}, \hat{U}, \hat{S} \approx \arg\max_{W,U,S} P(W, U, S|A) \tag{2.3}$$

Applying Bayes rule to the above Equation gives:

$$P(W, U, S|A) = \frac{P(A, S|U, W)P(U|W)P(W)}{P(A)} \tag{2.4}$$

Since $P(A)$ is constant for a given utterance and does not affect the outcome of the search, it is usually ignored. The remaining terms all constitute different components of the SUMMIT recognizer, which will be discussed in the sub-sections below.

2.2.2 Acoustic Model

The term $P(A, S|W, U)$ represents the probability of one specific segmentation and its associated acoustic observations, given the words and sub-word units. In this thesis, we will compute our acoustic observations given a particular segmentation S, to be described in more detail in Chapter 5. Given a particular segmentation S, $P(A, S|W, U)$ reduces to $P(A|S, W, U)$, which is known as the acoustic model. There are two main types of acoustic modeling approaches used in speech recognition, frame-based and segment-based.

Frame-based Modeling

In frame-based modeling, the acoustic observation space A, consists of a temporal sequence of acoustic frames (e.g. Mel-frequency ceptral coefficients or MFCCs) which

are computed at regular time intervals. To date, Hidden Markov Models (HMMs) have been the most dominant frame-based acoustic modeling technique for automatic speech recognition. However, HMMs have many limitations [22] which alternative models, such as segment-based models, have tried to address.

Segment-based Modeling

In segment-based modeling, frame-level feature vectors (e.g. MFCCs) are computed at regular time intervals, similar to the frame-based modeling approach. However, there is an additional processing stage in segment-based modeling which converts frame-level feature vectors to segmental feature vectors. SUMMIT hypothesizes phonetic landmarks at regions of large spectral change in the frame-level feature vectors. These variable-length landmarks are connected together to form a collection of possible segmentations of the speech utterance. For a given segmentation S, the acoustic observation space represents the feature vectors associated with S, as well as the feature vectors not associated with S, thus constituting the entire observation space [10].

2.2.3 Pronounciation/Lexical Model

$P(U|W)$ is the pronunciation or lexical model which gives the likelihood that a sequence of sub-word units U, was generated from a given word sequence W. This is achieved by a lexical lookup. Each word in the lexicon may have multiple pronunciations to account for phonetic variability [13].

2.2.4 Language Model

The language model is denoted by $P(W)$. $P(W)$ represents the *a priori* probability of a particular word sequence $W = \{w_1, w_2, \ldots, w_n\}$. SUMMIT typically uses an n-gram language model where the probability of each successive word depends only on the

previous $n-1$ words, as shown by Equation 2.5:

$$P(W) = P(w_1, w_2, \ldots, w_n) = \prod_{i=1}^{N} P(w_i | w_{i-1} \ldots w_{n-1}) \tag{2.5}$$

In this thesis, the language model $P(W)$ is an unweighted word-grammar pair, where a transition from one word to another can only occur if the word pair exists in at least one of the AV-TIMIT sentences [14].

2.2.5 Recognition Phase

Recognition in the SUMMIT system is accomplished by searching a weighted finite-state transducer (FST) [11], which is represented a cascade of smaller FSTs:

$$R = (S \circ A) \circ (C \circ P \circ L \circ G) \tag{2.6}$$

In Equation 2.6:

- S represents the acoustic segmentation described in Section 2.2.2
- A represents the acoustic observation space
- C relabels context-dependent acoustic model labels as context-independent phonetic labels
- P applies phonological rules mapping phonetic sequences to phoneme sequences
- L represents the lexicon which maps phoneme sequences to words
- G is the language model that assigns probabilities to word sequences

Intuitively, the composition of $(C \circ P \circ L \circ G)$ represents a pronunciation graph of all possible word sequences and their associated pronunciations. Similarly, the composition of $(S \circ A)$ is the acoustic segmentation graph representing all possible segmentations and acoustic model labelings of a speech signal. Finally, the composition of all terms in R represents an FST which takes acoustic feature vectors as

input and assigns a probabilistic score to hypothetical word sequences. The single best sentence is found by a Viterbi search through R. If n-best sentence hypotheses are needed, an A* search is then applied.

The search space consists of all possible segmentations of the acoustic features. In order to reduce the size of the search space and computation time of SUMMIT, an explicit segmentation phase is incorporated into the recognizer [18]. The segmentation phase will be discussed in more detail in Section 5.3.

Chapter 3

Sinusoidal Modeling of Speech

In this chapter we describe the acoustic theory behind speech production and the representation of this speech as a collection of sinusoidal components.

3.1 Acoustic Theory of Speech Production

The acoustics of speech production occurs in three distinct stages. First, a source of acoustic energy is created through interactions between airflow from the lungs and the laryngeal and supraglottal structures. Next the source is filtered by the resonant vocal tract cavities. Finally, speech is radiated from the lips [26].

Sources of sound in speech production are produced from three different sources, turbulence noise, vocal fold vibration and transients. Turbulence occurs due to rapid fluctuations in the velocity of airflow at a constriction. This causes the power spectrum of the noise source to be approximately flat. Turbulence noise can be produced at a constriction at the glottis, creating aspiration noise, or a constriction above the glottis, creating a frication noise. Unvoiced sounds are typically produced from turbulence noise.

Voiced sounds are produced from vocal fold vibration, which is caused by the opening and closing of the glottis. First, the glottis is closed off and pressure is built up behind the constriction. Eventually the pressure will build up and cause the vocal folds to push apart. The rapid airflow across the glottal opening then causes the

pressure to decrease, allowing the vocal folds to close and the cycle to repeat. This periodic opening and closing of the glottis is reflected in periodic glottal pressure and glottal volume velocity.

Transient sounds occur when there is a high pressure buildup behind a glottal constriction in the vocal tract. When the closure is opened, the pressure is suddenly released, resulting in a brief transient sound. Plosive bursts are an example of transient sources.

3.2 The McAulay-Quatieri Algorithm

The McAulay-Quatieri (MQ) Algorithm [19] is often used as a sinusoidal representation for sounds. The algorithm assumes that a speech waveform can be represented a collection of sinusoidal components of arbitrary amplitudes, frequencies and phases. For this thesis, we use a MATLAB implementation of the MQ Sinusoidal Model developed by Dan Ellis, an Associate Professor at Columbia [5]. First in the *analysis* stage, amplitude, phase and frequency parameters are extracted from the speech signal. Next in the *peak-matching* stage, tracks are formed among peaks which occur at similar frequencies. Finally, in the *synthesis* stage, extracted parameters are interpolated together to generate the synthesized speech output. These three sections, as well as extensions to the MQ model, will be discussed in more detail below.

3.2.1 Analysis

In the first step of the MQ Algorithm, amplitude, phase and frequency parameters are extracted from a speech waveform, as shown in Figure 3-1. To extract out these parameters, first the sampled speech waveform is broken down into a contiguous sequence of windowed frames each of length N, and an N-point short-time Fourier transform (STFT) of each frame is taken. Next, peaks in each STFT frame are found by determining locations where the slope of the waveform changes from positive to negative (concave down) and requiring peaks be at least a relative magnitude threshold below the largest peak in the frame. The MQ algorithm locates the amplitude and frequency

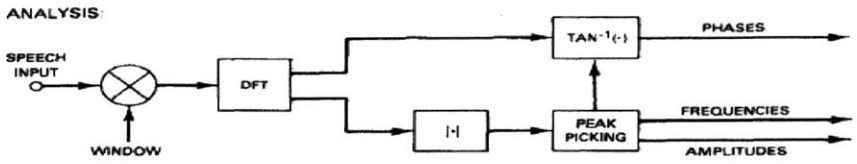

Figure 3-1: Block Diagram of Analysis Component [19]

for each peak found. The phase is determined by interpolating the unwrapped STFT phases to get exact peak phases for every sample point.

In this thesis, the speech signal was sampled at 16kHz. A 256-point STFT with a $16ms$ Hamming window was used. Finally, the speech utterance was analyzed at $4ms$ time intervals. These numbers were chosen to be similar to those used to obtain the MFCC acoustic feature vectors discussed in Section 2.2.2.

3.2.2 Peak-to-Peak Matching

As the fundamental frequency changes, the number of peaks from frame-to-frame changes. In particular, there is a rapid change in the number and location of peaks during voiced/unvoiced transitions. The concept of sinusoidal "births" and "deaths" is used to account for the movement of spectral peaks between frames. In order to match spectral peaks, tracks are formed by connecting peaks between contiguous frames. A new track is born if the frequency of a peak in the current frame is not within $\pm\Delta$ of the frequency of a peak in the previous frame. Similarly, a track is dead when there is no peak in the current frame that is within $\pm\Delta$ in frequency to a peak in the next frame. A magnitude increase threshold is also imposed so that contiguous peaks at the same frequency which have large magnitude differences are proposed to belong to different tracks. Figure 3-2 shows the birth and death of frequency tracks formed by connecting peaks of similar frequencies between frames.

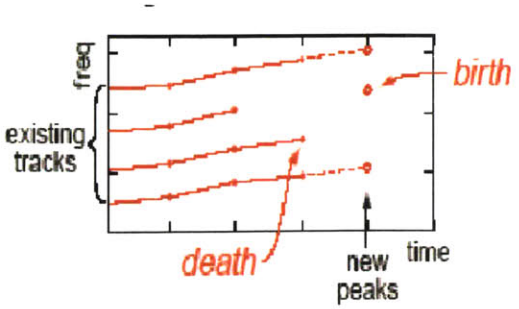

Figure 3-2: Birth and Death of Sinusoidal Tracks [4]

3.2.3 Synthesis

As shown in Figure 3-3, after reducing the speech waveform to a set of sinusoidal components, the MQ algorithm then synthesizes the waveform by interpolating the parameters of each track from frame to frame. The amplitude parameter is linearly interpolated between contiguous frames. However, the phase and frequency parameters cannot be linearly interpolated because these parameters are obtained modulo 2π. Thus the phase parameter must be unwrapped so that tracks are smooth and continuous across frame boundaries. In order to smooth out the phase between frames, the phase is interpolated with a cubic polynomial function, given by Equation 3.1:

$$\hat{\theta}(t) = \zeta + \gamma t + \alpha t^2 + \beta t^3 \tag{3.1}$$

To see the mathematics behind solving for the interpolation parameters, see [19]. Finally, after interpolation the sinusoidal tracks are added together to produce a synthetic speech output.

3.2.4 Improvements to Original MQ Algorithm

The original MQ Algorithm described above has some limitations [6] which newer, more improved models have tried to address. The sinusoidal model used in this thesis has been extended to address two of these problems.

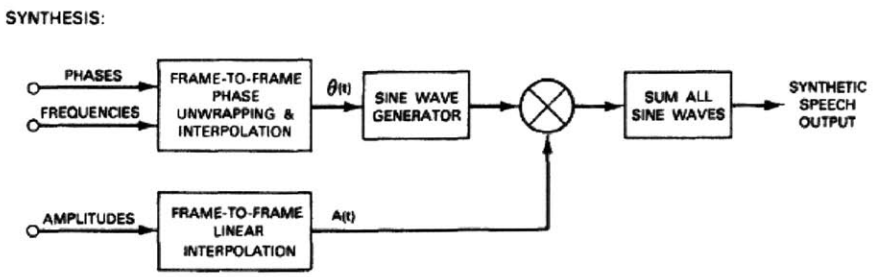

Figure 3-3: Block Diagram of Synthesis Component [19]

Lower Energy Threshold

As discussed in Section 3.2.1, peaks in the MQ algorithm are located by determining locations where the spectrum is concave down and at least a relative magnitude threshold below the largest peak in the frame. In regions of quiet speech where the spectrum is relatively flat, using a relative magnitude threshold results in many peaks being detected. Having many peaks detected causes a low amplitude hissing noise in the resynthesized waveform.

To prevent this added background noise, a constant absolute lower threshold is introduced. Thus the final threshold of a frame is the maximum of the largest relative threshold and the absolute lower threshold.

Hysteresis

In the original MQ model, tracks are observed to die out at a specific frequency and to be born again a few frames later at approximately the same frequency. These small tracks which die and appear again most probably belong to the same overall track, but have formed into separate tracks since the magnitude of the track has dropped below the relative magnitude threshold.

To allow tracks of similar frequency to combine into one track, a track amplitude hysteresis parameter is defined to be the lowest magnitude that a track may have without dying out. A larger hysteresis value means a lower magnitude before tracks end, thereby making tracks longer and smoother. Similarly, a smaller hysteresis value

results in smaller tracks.

3.3 Other Sinusoidal Modeling Techniques

Many other methods have been proposed for sinusoidal modeling of sound. In this section, we will discuss three such techniques.

3.3.1 Phase Vocoder

A phase vocoder is used to represent a speech signal as a series of sinusoidal components [8]. In the analysis stage, the speech signal is broken into windowed frames, and an N-point STFT of each frame is taken. Instead of extracting out the peaks from each SFTF frame as in the MQ algorithm, the phase vocoder considers all N frequency samples within a frame to be important. Therefore, each frame describes the evolution of a sound's frequency components over time. Thus, during the synthesis stage, the sound is reconstructed using all N frequency samples within each frame.

The phase vocoder model is appropriate for sounds with steady harmonic components. However, sounds with transient and noise components are not well represented using this model. In addition, the vocoder represents a sound at each frame by N time varying sinusoids, but not all these sinusoids are necessary to characterize a sound. For example, harmonic sounds can be modeled only by using sinusoids at integer multiples of a fundamental frequency. The MQ algorithm solves this problem by representing the sinusoids at each frame by the amplitudes and phases of the frequency peaks.

3.3.2 Spectral Modeling Synthesis

The Spectral Modeling Synthesis (SMS) Algorithm, developed at Stanford University, is yet another method for sinusoidal modeling [25]. The SMS method models an input sound *s(t)* as a sum of sinusoidal components and a noise component, as shown in

Equation 3.2:

$$s(t) = \sum_{i=1}^{k} A_k(t) \cos(\theta_k(t) + \psi_k) + e(t) \tag{3.2}$$

Similar to the MQ algorithm, the sinusoidal components are extracted by interpolating the spectral peaks. The residual noise is determined by subtracting the synthesized sinusoidal components from the original sound. This stochastic component is then modeled by filtered white noise, where the filter is determined by fitting a curve to the magnitude spectrum of the noise.

The MQ algorithm was originally used to model speech, but it showed promise for use on a broader class of sounds. However, sound often contains residual components and cannot be reduced to a small number of sine waves. Thus the SMS algorithm provides a more complete model for auditory signals.

3.3.3 Harmonic Plus Noise Model

The Harmonic Plus Noise Model (HNM) [28] models a speech signal *s(t)* as the sum of a harmonic plus noise component, as shown by Equation 3.3:

$$s(t) = s_h(t) + s_n(t) \tag{3.3}$$

The harmonic element represents the periodic components of a sound while the noise unit models the non-periodic components. The spectrum is divided into two bands. In the lower band, the signal is modeled by a collection of harmonically related sinusoids. The harmonic component is given by:

$$s_h(t) = \sum_{i=1}^{K(t)} A_k(t) \cos(k\theta(t) + \psi_k) \tag{3.4}$$

The signal in the upper band is assumed to be dominated by modulated noise. The noise part is modeled by convolving a time-varying autoregressive (AR) model $h(t, \tau)$ with Gaussian white noise $b(t)$ and then modulating the result by an energy envelope function $e(t)$. This noise part is given more explicitly by Equation 3.5:

$$s_n(t) = e(t)[h(t,\tau) * b(t)] \tag{3.5}$$

The HNM has been used to produce natural-sounding synthetic speech by applying different prosody and spectral envelope modification methods to both components. In addition, the HNM has also been used for smoothing diphone boundaries [27] and most notably in AT&T's Text-to-Speech System [1].

Chapter 4

Landmark Detection

Phonetic landmarks in speech utterances represent change from one phoneme to another and are usually identified by regions of large spectral change. In the sinusoidal model, rapid changes in spectral components are tracked using the concept of "birth" and "death" of the underlying sine waves. In order to determine exactly how to detect phonetic landmarks using the sinusoidal model, we look at how phonetic landmarks are placed with respect to the behavior of sinewave tracks generated from this model. The block diagram in Figure 4-1 shows the following steps in our landmark detector. Each of the following blocks will be discussed in the sections below.

4.1 Sinusoidal Model

The first step of the landmark detector is to analyze the speech signal using the sinusoidal model, discussed in Section 3.2. After this stage, we can model our speech signal $s[n]$ as a collection of sinusoidal components given by Equation 4.1:

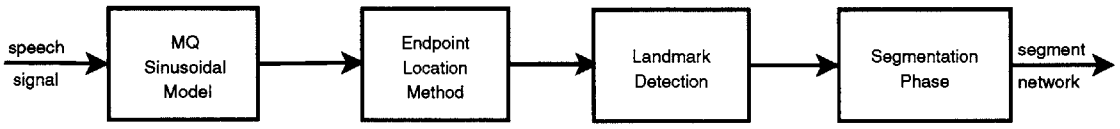

Figure 4-1: Block Diagram of Landmark Detector

$$s[n] = \sum_{k=1}^{N} A_k \cos(\theta_k[n] + \psi_k) \tag{4.1}$$

As discussed in Section 3.1, if the sound source is produced from vocal fold vibrations and is voiced, the spectrum will be periodic. A signal which is periodic in time is also periodic in frequency. Furthermore, any periodic signal can be represented by a collection of harmonically related sinusoids, known as the *Fourier series* [21]. Thus, voiced speech can be represented by a collection of harmonically related sinusoids.

If the sound source is produced by turbulence noise and is unvoiced, the spectrum will be flat and aperiodic. Similarly, the spectrum for transient sources is typically a short duration, intense energy spike. Turbulence and transient sounds can be modeled by a collection of sinusoids which are not harmonically related.

The sinusoidal components given by Equation 4.1 can be broken down into sinusoids which are harmonically related, representing voiced sound, and those which are not harmonically related, representing unvoiced speech. Equation 4.2 represents this decomposition:

$$s[n] = \underbrace{\sum_{k=1}^{N_1} A_k \cos(k\theta_k[n] + \psi_k)}_{voiced} + \underbrace{\sum_{k=1}^{N_2} A_k \cos(\theta_k[n] + \psi_k)}_{unvoiced} \tag{4.2}$$

Figure 4-2 shows typical sinusoidal tracks for voiced and unvoiced speech regions. Voiced sounds can be adequately estimated by a harmonic collection of sinusoids [23]. In voiced regions, peaks computed from the STFT of the waveform occur at close amplitudes and frequencies from frame to frame. Since tracks are connected by matching peaks at close frequencies in contiguous frames, the proximity in peak frequencies between frames results in long-duration tracks. In addition, the small amplitude and frequency variation of peaks results in slowly varying, smooth sinusoidal tracks.

According to the Karhunen-Loève analysis, unvoiced signals can only be sufficiently modeled by a very large number of sinusoids [29]. In unvoiced regions, peaks do not occur at close amplitudes or frequencies between neighboring frames. Here, the rapid frequency variation of peaks in unvoiced regions results in many short-duration

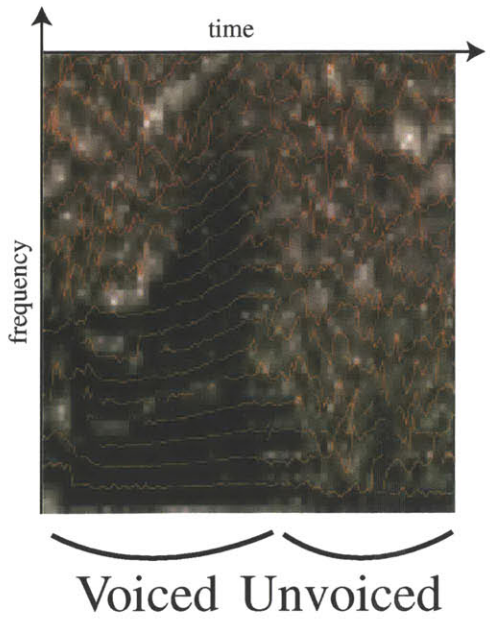

Figure 4-2: Typical sinusoidal tracks for voiced and unvoiced speech overlaid on a speech spectogram

tracks. Furthermore, the rapid amplitude and frequency variation of peaks causes the corresponding sinusoids to exhibit rapid fluctuations as well.

The births and deaths of the long, continuous sinusoids in voiced regions appear to occur at phoneme transitions. However, individual sinusoidal births and deaths can often occur too frequently and randomly to signal a phonetic transition. Therefore, in order to use the sinusoidal model to detect phonemes in voiced regions, voiced and unvoiced regions of the speech utterance must be detected.

4.2 Endpoint Location Method

An important problem in speech processing is to detect voiced speech in the presence of background noise, sometimes referred to as the endpoint location problem [24]. Accurately detecting the beginning and end of voiced speech segments will allow us to use the sinusoidal model to detect phonetic landmarks in these regions. This section will discuss our method of endpoint location.

4.2.1 Short-Time Energy

Short-time energy is often used in speech processing to distinguish between voiced and unvoiced speech segments [24]. The Short-time energy, *e[n]*, is defined to be the sum of the squared-magnitude of a windowed speech signal, i.e.:

$$e[n] = \sum_{m=1}^{N} (s[m]w[n-m])^2 \tag{4.3}$$

where *s[n]* are the speech samples and *w[n]* is the window size. In our method, *w[n]* is chosen to be 12ms.

After the speech utterance is passed through the sinusoidal model, the short-time energy is calculated on the synthesized output. Figure 4-3 shows a sinusoidal track representation for a signal and the corresponding short-time signal energy. When the signal energy is high, the sinusoidal tracks appear to be long and continuous, indicating regions of voicing. However, in unvoiced regions where sinusoidal tracks are short, the signal energy is usually very low. Voiced and unvoiced speech can often be differentiated by a corresponding low or high signal energy.

While the short-time energy or spectral energy has been conventionally used to distinguished between voiced and unvoiced segments, this measure becomes less reliable and robust in noisy environments [17]. Specifically, in noisy environments the signal energy is weak for particular phonemes, making it difficult to accurately detect regions of voicing. The phonemes for which it was was problematic to locate an exact endpoint include the following:

1. weak fricatives (/f, th, h/) at the beginning and end of a segment
2. weak plosive bursts for /p, t, k/
3. segment final nasals /n, m, η/
4. weak semivowels /r, y, l, w/

Therefore to obtain a more accurate estimation of voicing decisions, it is necessary to use an additional technique to identify voicing regions, which will be discussed in

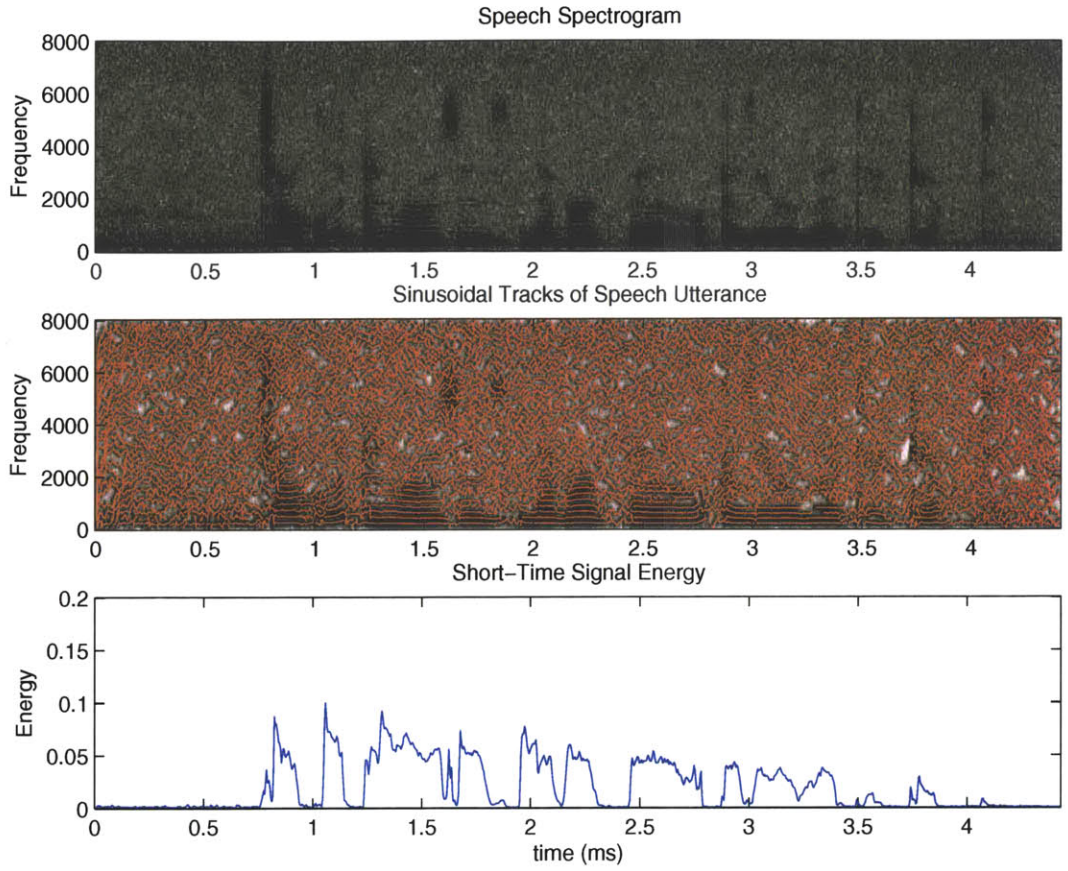

Figure 4-3: Sinusoidal Track Representation and Corresponding Signal Energy of Speech Utterance

the next section.

4.2.2 Harmonicity

Harmonicity is a measure of the strength of the pitch perception for a sound. Voiced regions can be modeled by a collection of harmonically related sinusoids, and thus contain high harmonicity. However, unvoiced regions are modeled with non-harmonic sinusoids and contain very little harmonicity. To exploit this difference, a harmonicity calculation is often used to detect regions of voiced and unvoiced speech segments. Harmonicity can be calculated as the the ratio of harmonic energy to signal energy, as given by Equation 4.4:

$$h[n] = \frac{s_h[n]^2}{s[n]^2}, \quad 0 < h[n] < 1 \tag{4.4}$$

To compute the energy of the harmonic signal, s_h, first the fundamental frequency of the synthesized waveform is calculated. We will discuss the method for fundamental frequency calculation in more detail in Section 4.3.1. The harmonic signal is calculated by finding the sinusoids which are integer multiples of the fundamental frequency.

Figure 4-4 shows a speech signal and its corresponding harmonicity. The harmonicity often peaks when transitioning into a voiced region since the harmonic energy is very large at the start of a voiced phrase. Similarly, the harmonicity tends to reach a minimum when entering an unvoiced region as the harmonic energy is very low.

Using harmonicity alone to identify regions of voicing has some disadvantanges. For example, as the SNR increases, the peaks and valleys of the harmonicity become more subdued. Figure 4-5 shows the harmonicity for the same speech signal shown in Figure 4-4, now corrupted by noise. The increased noise causes many small local peaks and valleys, which are often falsely detected to be voiced or unvoiced regions respectively.

However, whenever the harmonic energy is strong or weak, the harmonicity tends to peak or dip respectively. Especially in regions of weak voicing where the short-

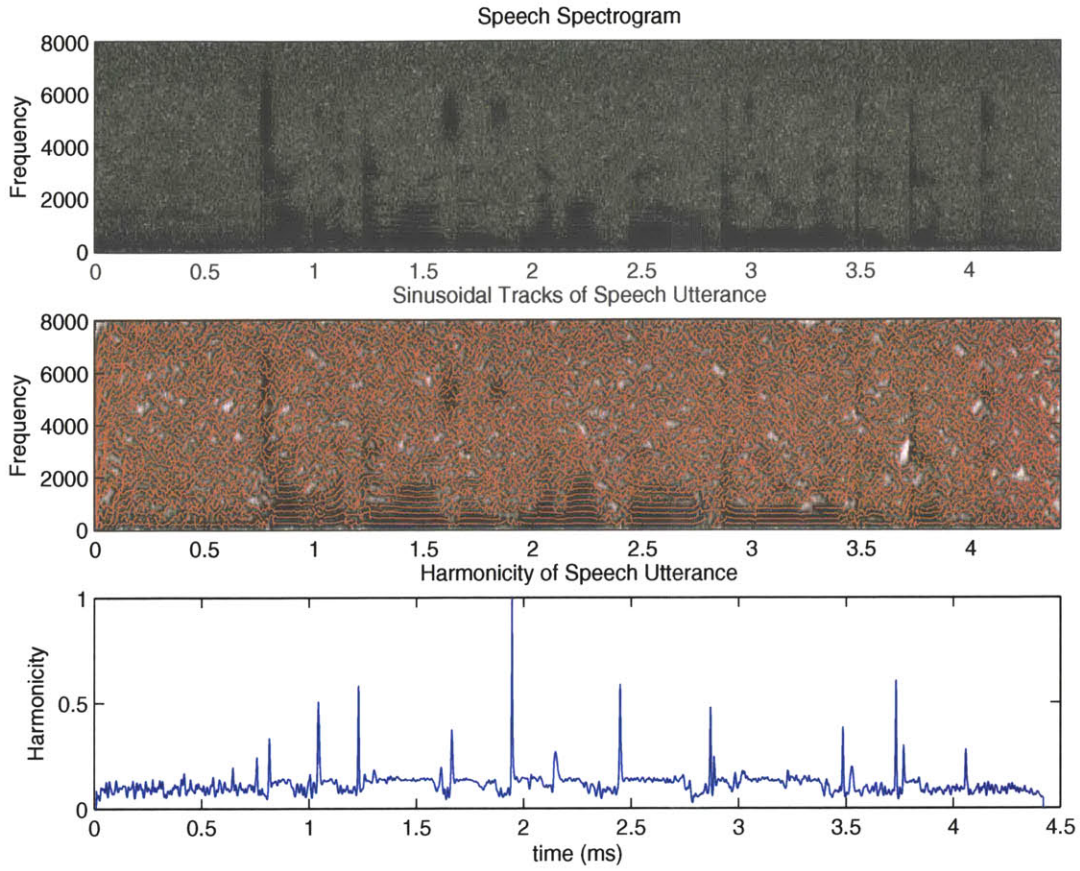

Figure 4-4: Sinusoidal Track Representation and Corresponding Harmonicity of Speech Utterance

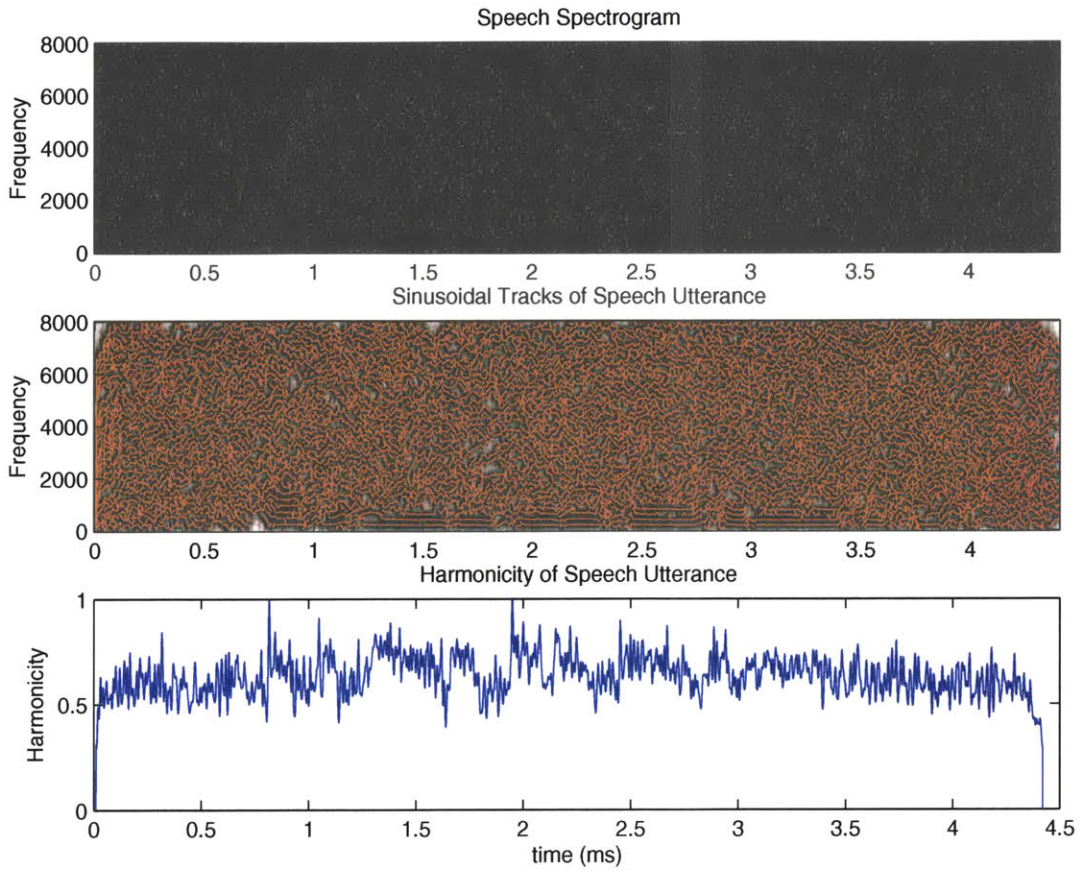

Figure 4-5: Sinusoidal Tracks and Harmonicity of Speech with a SNR of -5db. The speech is contaminated by white noise.

time energy does not provide a precise voicing decision, the harmonicity is more prominent and often helps to yield a more accurate voicing detection. Therefore, the signal energy may be used to identify general areas of voicing, but the harmonicity can help to make the areas of voicing more precise.

4.3 Detecting Landmarks from Sinusoidal Components

In order to determine exactly how to detect phonetic landmarks using the sinusoidal model, we look at how phonetic landmarks from AV-TIMIT waveforms are placed with respect to the behavior of sinewave tracks generated from the sinusoidal model. A few important observations are drawn by looking at the location of these landmarks with respect to the locations and behaviors of the sinusoidal tracks. These observations include:

- In Voiced Regions:
    1. Sinusoids tend to be long, smooth and slowly-varying
    2. A region with a lot of harmonically born sinusoids usually represents the beginning of a voiced region
    3. A region with a lot of harmonically dead sinusoids usually represents the transition from voiced to unvoiced region.
    4. Oftentimes sinusoidal births and deaths are not present when transitioning into a semivowel or nasal, or when transitioning from one vowel into another
- In Unvoiced Regions:
    1. Sinusoids tend to be short and rapidly-varying
    2. Births and deaths of sinusoids occur frequently and randomly with respect to AV-TIMIT phonetic landmarks

With these observations in mind, the method for detecting phonetic landmarks from sinusoidal components in voiced and unvoiced regions will be described in the following sections.

4.3.1 Identifying Harmonically Related Sinusoids

In voiced regions, phonetic landmarks are detected from the births and deaths of sinusoids. Sometimes there are not breaks in sinusoidal tracks when transitioning between particular voiced phonemes, but there is usually a break in harmonically related sinusoids. The births and deaths of harmonically related sinusoids, which are obtained from knowledge of the fundamental frequency, allow us to more accurately detect landmarks rather than just the sinusoidal tracks alone. In this section, we will describe the method for obtaining harmonic sinusoids.

Fundamental Frequency

In order to detect harmonic sinusoidal components, it is first necessary to identify regions where sinusoids are harmonically related to the fundamental frequency. One way of estimating the fundamental frequency of a speech segment is to compute the cepstrum. The cepstrum is a Fourier analysis of the logarithmic amplitude spectrum of the signal. If the log amplitude spectrum contains many regularly spaced harmonics, then the cepstrum will show a peak corresponding to the fundamental frequency of these harmonics.

To obtain the fundamental frequency of a speech signal, the waveform is broken into frames and the fundamental frequency for each frame is computed by finding the peak in the cepstrum of the frame. Accurately detecting the pitch period of a speech signal is difficult for several reasons [3]. For example, signals are not always completely periodic and can be corrupted by noise. Therefore, we take the most dominant peak of the cepstrum in each frame to calculate the fundamental frequency.

Harmonically Related Sinusoids

After the fundamental frequency for the waveform is computed, sinusoidal tracks which occur at frequencies that are multiples of the pitch are identified to be harmonically related sinusoids. As shown in Figure 4-3, the higher energy speech harmonics at low frequencies are less corrupted by noise and their long, smooth behavior tends to indicate phonetic transitions. However, high frequency harmonics are typically lower in energy and are more corrupted by noise, therefore providing little information about phonetic transitions. Therefore for the purposes of detecting landmarks from sinusoidal components, harmonic sinusoids are only detected at low frequency regions, up to 4 kHz.

4.3.2 Landmarks from Harmonic Sinusoids

After harmonically related sinusoids are identified in regions of voicing, the next step is to detect phonetic landmarks from the births and deaths of these tracks. First, the number of sinusoids that are born or die at a set frame interval of 12ms is counted. However not all sinusoidal births and deaths correspond to potential phonetic landmarks. For example, when transitioning into the beginning of a voiced region, various sinusoids may not be born at the exact same instance, but rather there is a sequence of staggered births. Furthermore, noise may influence sinusoidal tracks and cause them to break at certain frames but this usually does not correspond to a phonetic landmark. Therefore, the optimal number of harmonic births which constitute a potential phonetic landmark must be determined. In addition, the optimal number of harmonic deaths to identify a potential phonetic landmark must also be calculated. We will discuss our method for determining these optimal numbers in Section 6.2.1.

In addition, we further make the assumption that if a harmonically born landmark is detected, another harmonically born landmark will not be detected for at least a certain number of frames future frames. This parameter will be referred to as born hop. Similarly, if a harmonically dead landmark is detected, another similar landmark will not be detected for at least a specified of frames, known as dead hop. Again, we

will discuss our method for optimally determining these parameters in Section 6.2.1.

Finally, sinusoidal births and deaths can oftentimes occur close to each other when transitioning between phonemes. For example, the movement between two phonemes is sometimes characterized by the birth and death of many harmonic sinusoids. In order to remove landmarks which essentially detect the onset of the same phoneme, we search for locations where birth and death landmarks are within a window length and remove all but one landmark from this window. This window length will be referred to as the harmonic window.

4.3.3 Landmarks in Unvoiced Regions

As stated above, in unvoiced regions sinusoids tend to be short and rapidly varying, and thus the births and deaths occur too frequently to indicate phonetic transitions. Furthermore, the full segmentation approach, which places potential landmarks every 30ms, has a much lower word error rate in the presence of noise than the original segmentation approach. Due to the lack of information from the sinusoidal model in unvoiced regions and the lower word error rate of the full segmentation, we decided to place landmarks at least every 30ms apart. We also found that long unvoiced regions contain very few phonemes, and thus placing landmarks greater than 30ms apart was enough to detect these phonemes. The following outlines our criterion for placing landmarks in unvoiced regions:

- If unvoiced region is less than 75ms, place landmarks every 28ms[1]
- If unvoiced region is less than 300ms, place landmarks every 40ms
- If unvoiced region is greater than 300ms, place landmarks every 64ms

After landmarks are detected, these landmarks are interconnected together to form a network of hypothetical segmentations. Our next step is incorporate an explicit segmentation stage into this network in order to reduce the size of the search space, the main topic of the next chapter.

[1]The resolution of the sinusoidal model results in landmarks must be placed at time intervals which are multiples of 4. Thus landmarks were placed at 28ms intervals instead of 30ms intervals.

Chapter 5

Segmentation

Segmentation is incorporated prior to the search phase in order to decrease recognition computation time. In this chapter, we discuss different segmentation approaches for the full segmentation method, original segmentation method and sinusoidal model.

5.1 Background

As discussed in Section 2.2, phonetic landmarks specify a collection of possible segmentations for the utterance. It is computationally expensive to search through this large segmentation network. Therefore, an explicit segmentation phase is incorporated into the recognizer to reduce the size of the search space and the computation time of the recognizer.

In the segmentation phase, the segment graph is pruned prior to search. However this pruning also increases the number of deletion errors. Therefore, in creating segment networks, there is a tradeoff between minimizing search space and minimizing errors.

In the following sections, we will discuss the different segmentation approaches used for the full segmentation and original segmentation methods, as well as and our sinusoidal model segmentation. We save the results of these different approaches for Chapter 6.

5.2 Full Segmentation

Figure 5-1 illustrates interconnections among landmarks in the full segmentation method. In this approach, hypothetical landmarks are placed at fixed spacings, independent of spectral content. The segment graph is not pruned prior to search. Therefore, the segmentation network is very large, consisting of all possible landmark interconnections within a maximum segment length. In this thesis, the full segmentation approach uses a 30ms fixed landmark rate and a maximum segment length of 250ms. Figure 5-2 shows the graphical display of the segment network from SUMMIT for the full segmentation approach. As we will observe in Chapter 6, the computation time for the full segmentation method is very large.

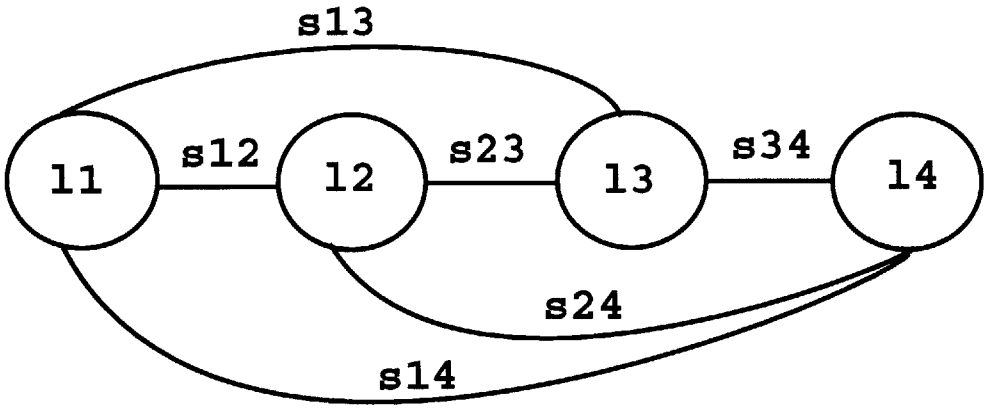

Figure 5-1: Segment network for Full Segmentation technique. Each landmark l_i is fully connected to every other landmark l_j in the graph via segments s_{ij}.

5.3 Original Segmentation

As discussed in Section 2.2.2, landmarks in the original segmentation method are hypothesized at regions of large spectral change within the MFCC frame-level feature vectors. More specifically, major landmarks are hypothesized at locations where

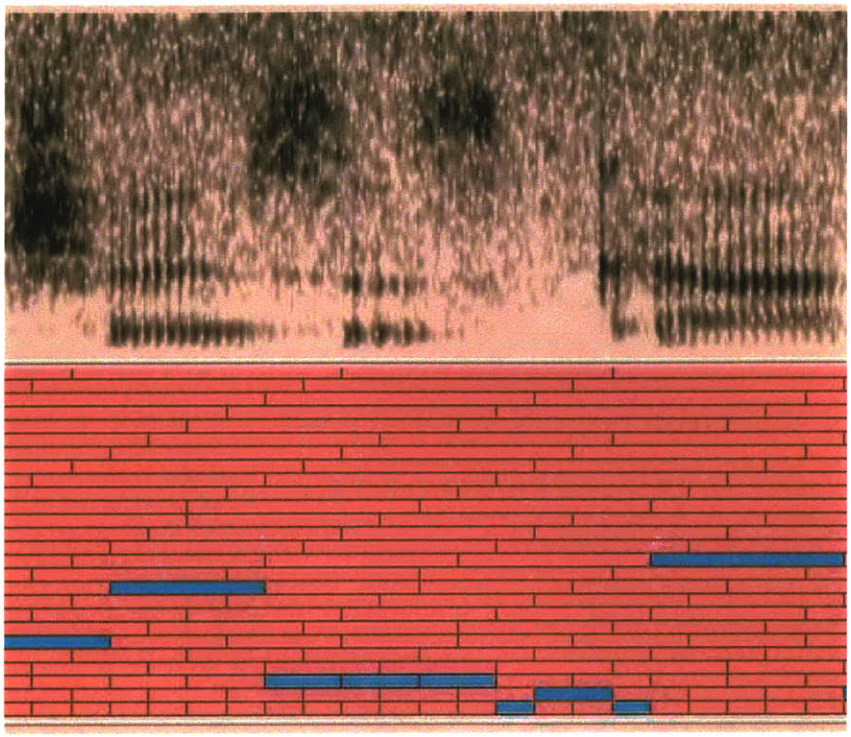

Figure 5-2: Graphical display of the segment network for the full segmentation approach from the SUMMIT recognizer

the spectral change exceeds a specified global threshold. A fixed density of minor landmarks are detected between major landmarks where the spectral change, based on the fixed minor landmark density, exceeds a specified local threshold. All minor landmarks are fully interconnected between, but not across major landmarks, to form a segment network. In addition, each major landmark is connected to two major landmarks forward. Figure 5-3 shows a typical segment network formed from major and minor landmarks, and Figure 5-4 illustrates the corresponding graphical display from SUMMIT.

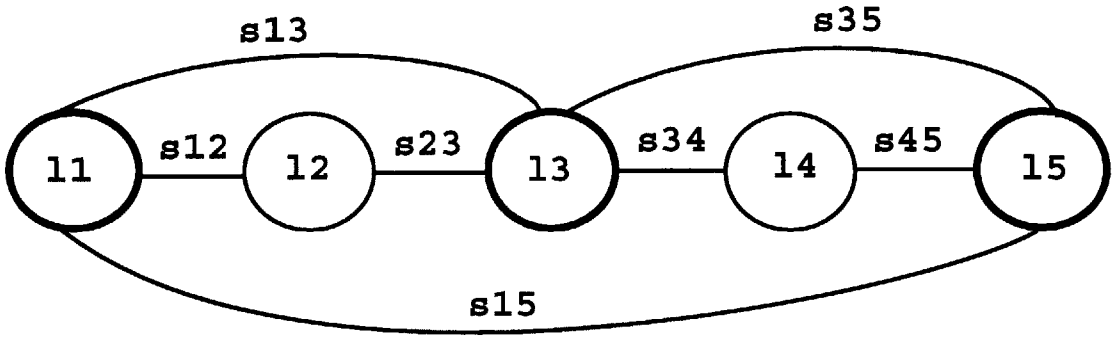

Figure 5-3: Segment network for Original Segmentation technique. Major landmarks are indicated in bold. Each minor landmark l_i between major landmarks is fully connected to every other landmark l_j in the graph via segments s_{ij}. In addition, each major landmark is connected to two major landmarks forward.

5.4 Sinusoidal Model Segmentation

In Chapter 4, we discussed our method for detecting landmarks using the sinusoidal model. The placement of these landmarks is determined by the behavior of sinusoidal tracks. Sinusoidal landmarks which are not hypothesized to be major landmarks are termed minor landmarks. Major landmarks serve as hard boundaries for interconnections among minor landmarks. Because of the tradeoffs associated with segmentation, we explore a variety of segmentation methods in this thesis.

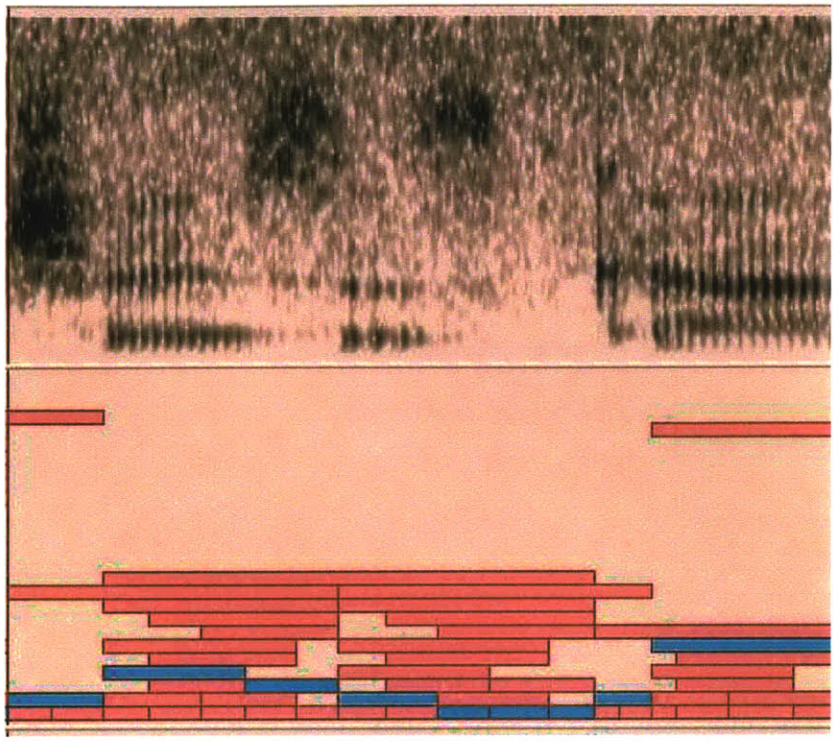

Figure 5-4: Graphical display of the segment network for the original segmentation approach from the SUMMIT recognizer

5.4.1 Segmentation at Voiced/Unvoiced Boundaries

Section 4.2 discusses two methods for determining voiced and unvoiced decision boundaries, using a short-time energy measurement as well as a combined short-time energy and harmonicity measurement. The most obvious choice for major landmarks are at these decision boundaries where there are many sinusoidal births and deaths. In the middle of voiced regions, there are no obvious trends observed in sinusoidal behavior which indicates placing a major landmark. As mentioned in Section 4.3.3, the short and somewhat random behavior of sinusoidal tracks in unvoiced, burst and non-speech regions also provides little indication of placing major landmarks within this region. In this chapter, we discuss placing landmarks when voiced boundaries are detected using a short-time energy measurement, as well as using the combined measurements.

Segmentation Using Short-Time Energy

One method we explored was placing major landmark at voiced and unvoiced decision boundaries detected by the short-time energy measurement. As mentioned in Section 4.2.1, there are certain phoneme types which are difficult to locate an exact endpoint using the energy measurement. Our recognizer can be somewhat sensitive to the placement of major landmarks, as small movements of the landmark locations can results in completely different word hypothesis. Therefore, while short-time energy can provide a crude estimate of major landmarks, the necessity for precise landmark locations makes it important to incorporate other voicing decision measures.

Segmentation Using Short-Time Energy and Harmonicity

As discussed in Section 4.2.2, a harmonicity measurement was used in addition to short-time energy to help make voicing decisions more precise. As we show in Chapter 6, adding in the harmonicity measurement gives more accurate decisions about placing major landmarks, particularly in the problematic phoneme areas. These small changes in major landmark locations lead to large improvements in word accuracy.

5.4.2 Segment Connectivity Methods

In addition to determining location for major landmarks, we also explored different connectivity methods between the segments. In this section, we will discuss three such methods.

One-Connection

In the one-connection method, all minor landmarks between major landmarks are fully interconnected. In addition, each major landmark is connected to the next two consecutive major landmarks. This is the same connectivity approach used in the original segmentation method, and is shown in Figure 5-3.

Two-Connection

Even with using short-term energy and harmonicity measurements to determine major landmarks, there are some major landmarks which are not placed exactly between voiced and unvoiced separations. Particularly, as the signal-to-noise ratio decreases, the signal power gets weaker and it becomes more difficult to accurately detect endpoint locations. As a result, major landmarks may be placed in the middle of actual phonemes.

To increase segment interconnections across major landmarks and correct for major landmarks placed at low signal energies, we introduce a two-connection method. In this method, each minor landmarks is connected to landmarks which fall up to two major landmarks away. In addition, each major landmark is connected to the next three consecutive major landmarks. This connnectivity technique is shown in Figure 5-5.

Partial-Connection

In the second-connection method, segments between two major landmarks are joined in hopes of improving connectivity across major landmarks placed at low signal energy levels. However, some major landmarks which occur across a large signal energy

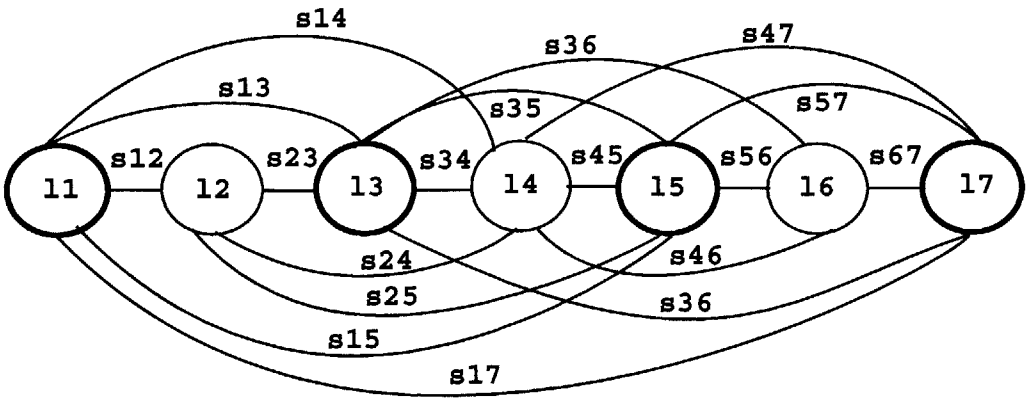

Figure 5-5: Segment network for Sinusoidal Model Two-Connection technique. Major landmarks are indicated in bold. Each minor landmark l_i is connected to every other landmark l_j which falls up to two major landmarks away via segments s_{ij}. In addition, each major landmark is connected to the next three major landmarks.

difference are placed at very accurate voiced and unvoiced decision regions. A two connection method across these landmarks increases the search space but has very little effect on improving the recognizer performance.

To decrease segment interconnections across major landmarks when necessary, we label major landmarks based on the signal energy difference on either side of the landmark. Landmarks which have a signal energy difference above a specified threshold are defined to be hard landmarks, while soft landmarks have an energy difference below this threshold. Minor landmarks can be connected to other minor landmarks across soft major landmarks. However, minor landmarks cannot be connected across hard major landmarks. In addition, each major landmark is connected to the next three consecutive major landmarks. Figure 5-6 illustrates the connectivity using hard and soft major landmarks more explicitly. Notice the smaller number of connections in this figure compared to the two-connection method in 5-5. The graphical display of the partial-connection method from SUMMIT is shown in Figure 5-7.

5.4.3 Segmentation Using MFCC Distance Information

Our method of placing major landmarks at regions of large energy change has some disadvantages. For example, if the utterance contains very quick, short noise clicks or

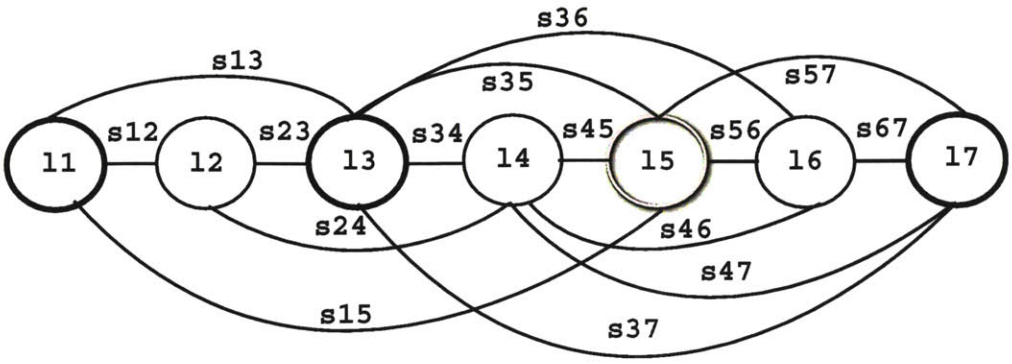

Figure 5-6: Segment network for Partial-Connection technique. Hard major landmarks are indicated in bold while soft major landmarks are embossed. Minor landmarks l_i are connected across a soft major landmarks to other landmarks l_j which falls up to two major landmarks away via segments s_{ij}. However, minor landmarks cannot be connected across hard major landmarks. In addition, each major landmark is connected to the next three major landmarks.

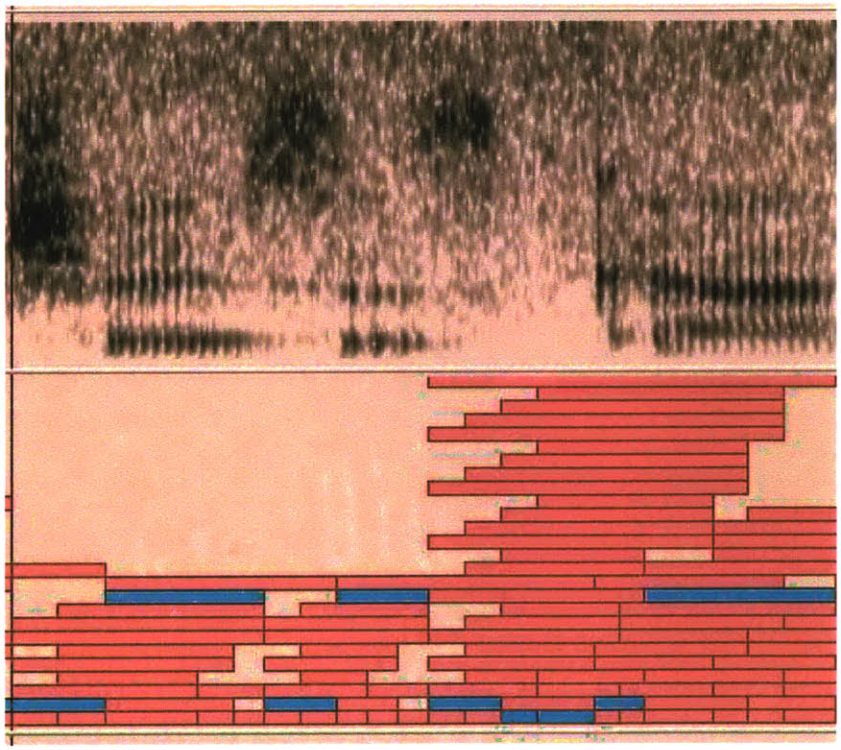

Figure 5-7: Graphical display of the segment network for the partial-connection method from the SUMMIT recognizer

bursts, major landmarks are hypothesized due to the energy difference. To correct for these possible errors, we also explore looking at information from MFCC distances.

The original segmentation algorithm detects landmarks based on large spectral changes in MFCC frame-level feature vectors. The 14-dimension MFCC feature vector for frame i is given by:

$$\vec{m}_i = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{14} \end{bmatrix} \tag{5.1}$$

To compute the MFCC distance between frames i and j, we calculate the sum of squared difference between each of entries m_i and m_j and then take the square root. This distance d_{ij} is given by Equation 5.2:

$$d_{ij} = \sqrt{\sum_{n=1}^{14} [\vec{m}_i(n) - \vec{m}_j(n)]^2} \tag{5.2}$$

This distance is computed between all possible frames of the utterance to form an MFCC distance matrix. Equation 5.3 shows this distance matrix for an signal of n frames and a plot of this matrix is illustrated in Figure 5-8.

$$\mathbf{D} = \begin{bmatrix} d_{11} & d_{12} & \dots & d_{1n} \\ d_{21} & d_{22} & & \\ \vdots & & \ddots & \\ d_{n1} & & & d_{nn} \end{bmatrix} \tag{5.3}$$

After the distance matrix is calculated, we observe the behavior of sinusoidal major landmarks with respect to distances between feature-vector frames. Major landmarks should be placed at locations of a large difference in spectral content on either side of the landmark, as shown in Figure 5-9. However, some major landmarks hypothesized by the sinusoidal model approach occur with little change in MFCC distance, most likely due to quick bursts of energy. Figure 5-10 illustrates a major

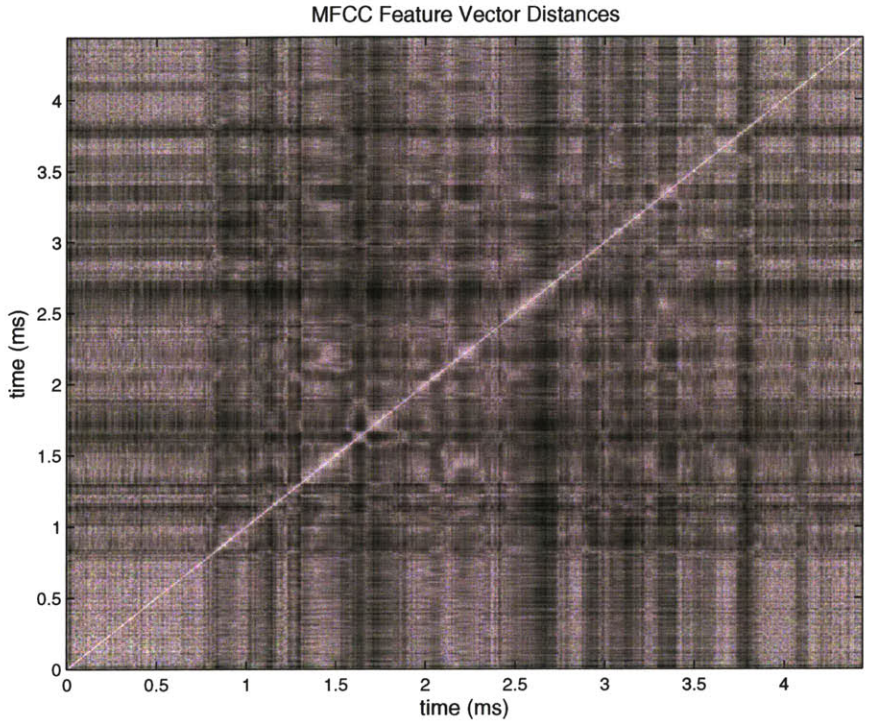

Figure 5-8: MFCC feature vector distance matrix of a speech signal.

landmark with little change in MFCC distance. Therefore, we look at recognition peformance when we remove major landmarks with little difference in MFCC distance between the previous and following frames. In order to remove these possible major landmarks, we calculate an average windowed MFCC distance on either side of each major landmark. If the difference between the two distances is less than a specified threshold, we remove this major landmark.

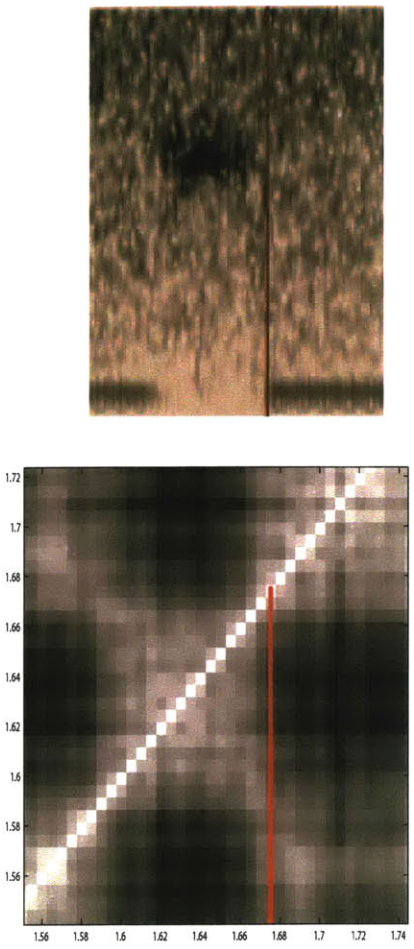

Figure 5-9: The top figure shows a major landmark overlaid on top of the speech signal. Notice the large change in MFCC distance on either side of the major landmark, as illustrated in by the bottom figure.

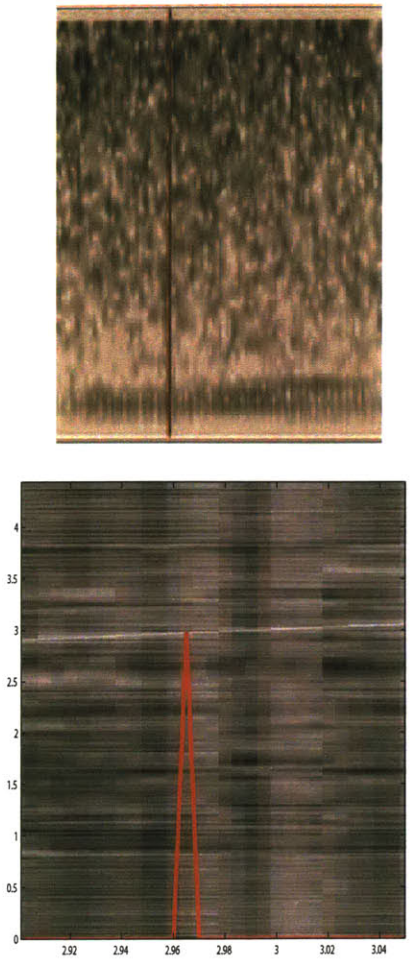

Figure 5-10: The top figure shows a major landmark overlaid on top of the speech signal. The bottom figure illustrates a small change in MFCC distance on either side of the major landmark.

Chapter 6

Experimental Results

In this chapter we look at the performance of the sinusoidal model on the noisy AV-TIMIT sentences. Specifically, we analyze the landmark detection and segmentation algorithms proposed. Finally, we compare the performance of the sinusoidal model-based algorithm with the original segmentation and full segmentation methods under three different noise conditions.

6.1 Experimental Setup

Our recognition experiments draw from the AV-TIMIT corpus. The vocabulary for this corpus consists of 1793 words. As mentioned in Section 2.2.4, the language model used in this thesis is an unweighted word-pair grammar, where a transition from one word to another can only occur if the word pair exists in at least one of the AV-TIMIT sentences. Because 1411 words in the corpus occur in only one of the 453 AV-TIMIT sentences, this heavily constrains the grammar [14].

In this thesis, the recognizer is evaluated using a word recognition paradigm rather than a phonetic recognition paradigm which is also commonly used. Many different experiments comparing the word error rate and recognizer computation times of different models are discussed in this chapter. Each model is tested on noisy AV-TIMIT sentences from the test set, unless otherwise specified. We discuss results for three different noise types from the Noisex-92 database, namely white noise, babble noise

and destroyer operations room noise. The SNRs for the noisy speech are varied from -10db to 20db in 5db increments. Finally, each model is tested using acoustic models specific to the SNR and noise type.

6.2 Sinusoidal Model Landmarks

Landmarks are detected from sinusoidal "births" and "deaths" and represent a boundary or transition from one phoneme to another. In this section, we will discuss different experiments and conclusions relating to these sinusoidal model landmarks.

6.2.1 Landmark Detection Parameter Settings

In Section 4.3, we discussed our method for detecting landmarks from sinusoidal components. In order to appropriately place these landmarks, the values for five parameters in our landmark detection method must be determined. These parameters include the number of harmonic births, number of harmonic deaths, born hop, dead hop and harmonic window.

To find these values, we look at how hypothetical landmarks are placed in the sinusoidal model compared to the actual phonetic boundaries obtained from forced transcriptions of clean speech. A certain parameter setting that correctly predicts landmarks at all the phonetic boundaries for a specific waveform might perform poorly on another waveform. In addition, a parameter setting that predicts landmarks at all phonetic boundaries may also overgenerate too many hypothetical landmarks. Therefore, we want a setting which finds a balance between detecting phonetic boundaries and overgenerating landmarks.

A receiver operating characteristic (ROC) curve is a common tool used to find a suitable tradeoff between detection and overgeneration as parameter settings are varied. An ROC curve, shown in Figure 6-1, plots the probability of detection, P_d, versus the probability of false alarm, P_{fa}. The probability of detection is defined as the total number of phonetic boundaries that are correctly detected by our landmark detection method divided by the total number of phonetic boundaries. The probability of false

alarm, P_{fa} is defined as the total number of over-generated landmarks divided by the total number of frames in a speech signal where landmarks could potentially be placed.

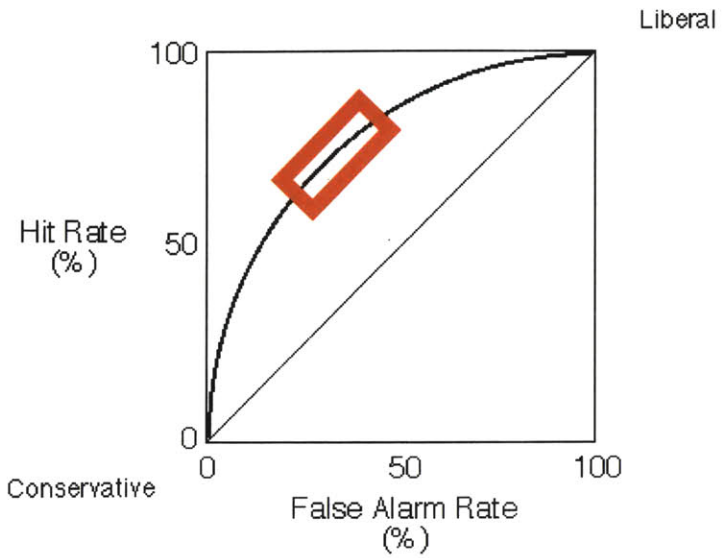

Figure 6-1: Example Receiver Operating Characteristic (ROC) Curve plotting the Probability of Detection, P_d versus the Probability of False Alarm, P_{fa}

In order to determine an appropriate setting of the parameters, we vary each parameter as follows:

- number of harmonic births: varied from 1-7 frames in one step increments
- number of harmonic deaths: varied from 1-7 frames in one step increments
- born hop: varied from 12-48 ms in 4ms increments
- dead hop: varied from 3-48 ms in 4ms increments
- harmonic window: varied from 0-28ms in 4ms increments

For each waveform, the probability of detection and the probability of false alarm are calculated for each setting of the parameters. (P_d, P_{fa}) pairs are computed for 50 AV-TIMIT waveforms for each signal-to-noise ratio level, and the (P_d, P_{fa}) pairs are

then averaged for each parameter setting. After calculating average (P_d, P_{fa}) pairs over all waveforms, an ROC curve is generated, as illustrated in Figure 6-2. (P_d, P_{fa}) values represent a tradeoff - if we detect more AV-TIMIT landmarks we increase P_d at the cost of P_{fa}, whereas if we detect less of the AV-TIMIT landmarks, the opposite effect will occur. The optimal spot on the ROC curve is for parameter settings that generate a high P_d but a low P_{fa}, as indicated by the rectangle in Figure 6-1.

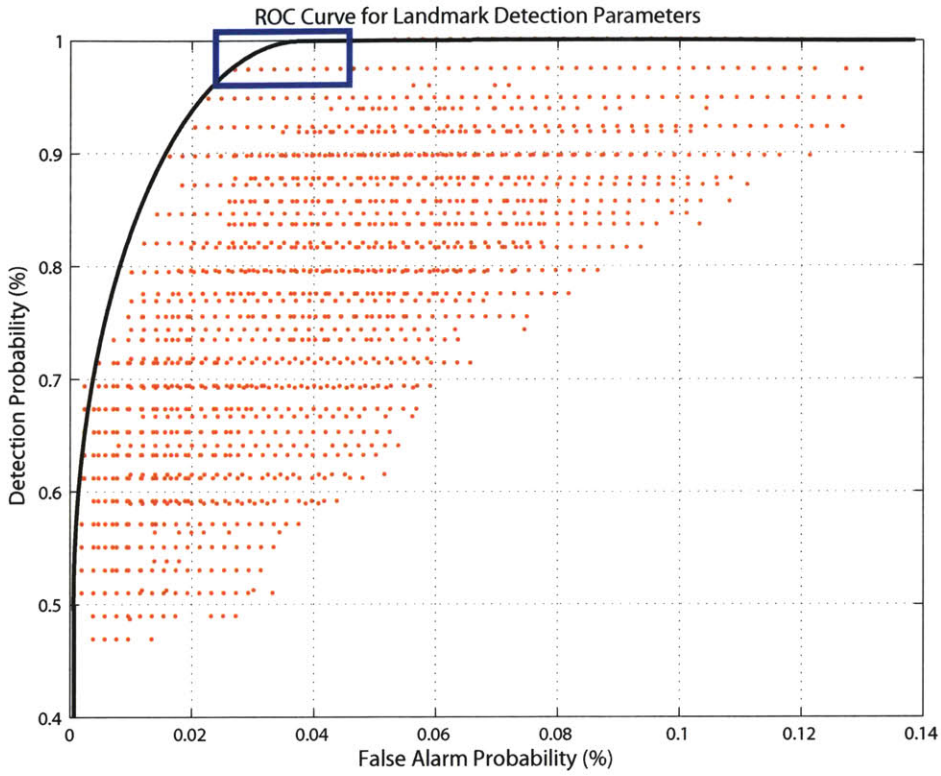

Figure 6-2: Generated Receiver Operating Characteristic (ROC) Curve for Landmark Detection Parameters

We found a range of parameter settings within this optimal region on the ROC curve, shown by the rectangle in Figure 6-2. Each dot represents average (P_d, P_{fa}) values for a specific setting of the parameters. Similarly, the curve represents the optimal (P_d, P_{fa}) values as the parameter settings are varied. It is important to stress that the optimal parameters were determined based on waveforms in the development set. After looking at the parameter values and further observing the placement of

these landmarks with respect to the onset of AV-TIMIT phonemes, we determined a setting of values which appeared to be reasonably located at actual phoneme onsets. The final parameter values used in this thesis to detect phonetic landmarks include:

- number of harmonic births: 2
- number of harmonic deaths: 2
- born hop: 24 ms
- dead hop: 28 ms
- harmonic window: 20 ms

6.2.2 Phonetic Detection Probability

To analyze the accuracy of the sinusoidal model in detecting phonetic landmarks, the placement of these landmarks is compared against the transcribed phonemes as estimated from the forced alignments of clean AV-TIMIT utterances. In this experiment, if the sinusoidal landmark is within 20ms of the phonetic boundary in the forced alignment, we say that the boundary is detected. The detection probability of the onset of the different AV-TIMIT phonemes [1] was computed using all 285 waveforms in the AV-TIMIT development set. Figure 6-3 shows the detection probability for the onsets of the different AV-TIMIT phonemes in the clean speech condition.

As expected, the onsets of semivowels, such as $/r, y, l, w/$ have a low detection probability. Sometimes there are not breaks in harmonic sinusoidal tracks when transitioning from a vowel into a semivowels. Furthermore, semivowels are often characterized by weak signal energy at the onset of the phoneme, resulting in tracks not hypothesized at the onset of these semivowel.

In addition, nasals onsets such as $/m, m, \eta/$ also have a low detection probability. These phonemes are weak and prominent only at low frequencies making it difficult to precisely identify sinusoidal births under noisy conditions.

[1] For a listing of the AV-TIMIT phonemes, please see Appendix A.

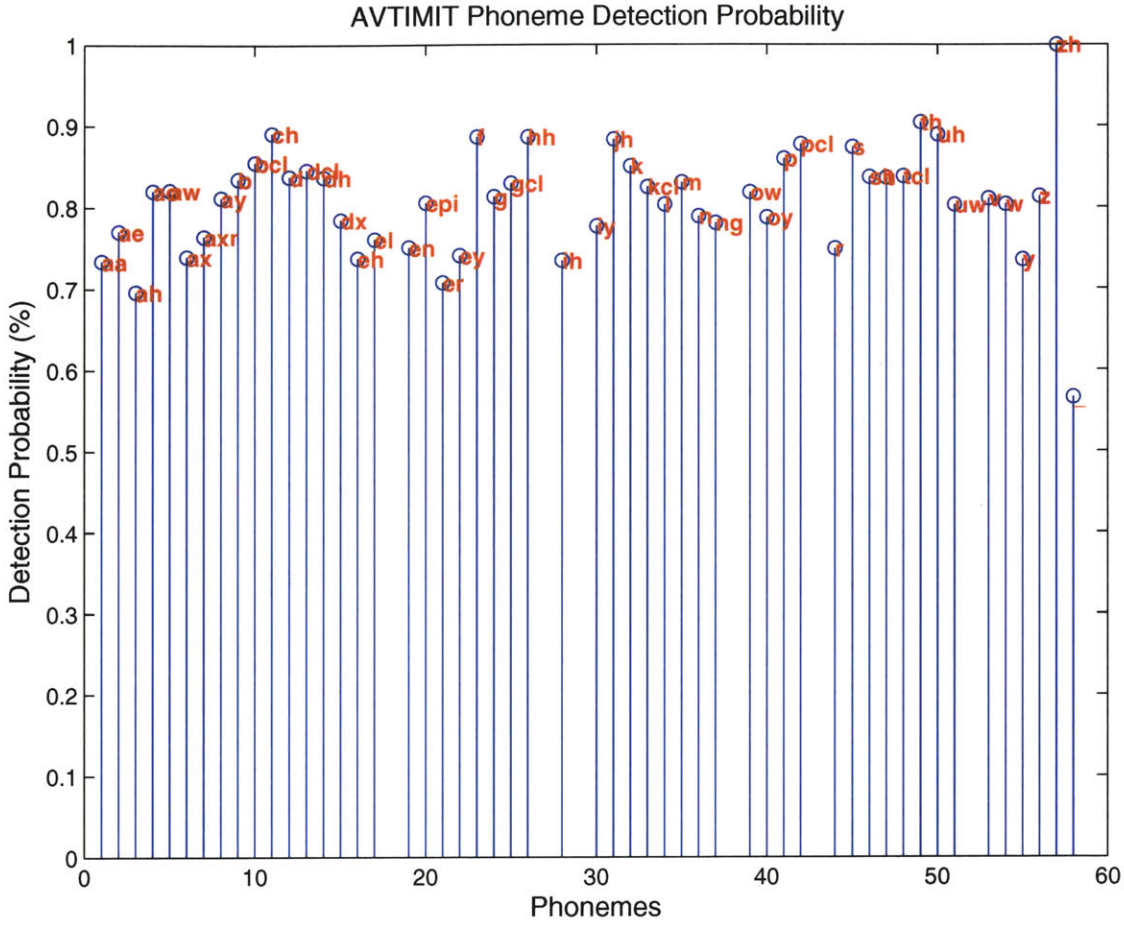

Figure 6-3: AV-TIMIT phoneme detection probability under clean speech

The onsets of unvoiced phonemes tend to have a slightly larger detection probability than voiced phonemes. Landmarks are placed at least every 30ms in unvoiced regions. This placement is generally more frequent than landmarks placed in voiced regions, resulting in a higher detection probability of these unvoiced phonemes.

6.2.3 Landmark Types Not Used

It seems natural to add landmarks based on large spectral change and near areas of consonant bursts. However, we found that neither of these landmark types helped to improve our landmark detector.

Energy Landmarks

In the SUMMIT system, hypothetical landmarks are placed at regions of large spectral change. To see if our system would benefit from adding landmarks at these regions, we briefly looked at placing landmarks at large changes in sinusoidal energy and found no improvement in word error rate.

When a speech signal is corrupted by noise, energy landmarks are sometimes placed when there are large noise changes in the signal. However, these hypothetical energy landmarks often occur close to many sinusoidal births and deaths, and do not help to provide any additional information about hypothetical phonemes. Since, energy landmarks did not help to improve the word error rate, they were not used in the landmark detector.

Burst Landmarks

Often times if a consonant burst comes before a region of voicing, it occurs so weakly that it is hard to identify this burst landmark. Many of the undetected consonant bursts in our model occur in unvoiced regions right before the birth of many sinusoidal tracks into a voiced region. We noticed that adding landmarks a few milliseconds before the onset of sinusoidal births to detect these phoneme bursts (i.e. burst landmarks) decreases the word error rate for clean speech. However, under noisy

conditions the recognizer is sensitive to landmarks which are close together. Particularly because these burst landmarks are placed in unvoiced regions which is often corrupted by noise, the recognizer often incorrectly hypothesizes these landmarks to be the onset of fricatives or burst phonemes. Hypothesizing one phoneme wrong can cause an entire word or sentence to be hypothesized incorrectly. Therefore, adding in these burst landmarks also did not improve the error rate, particularly as the noise level increased.

6.3 Sinusoidal Model Segmentation Methods

In Section 5.4 we discussed various sinusoidal model segmentation approaches we investigated. This included using various voicing decision methods, different segment connectivity approaches and added MFCC distance information. In this section, we compare word error rate and computation time results for the following segmentation techniques:

- Energy Voicing Decision, One-Connection Method - *eng+onecnct*
- Energy and Harmonicity Voicing Decision, One-Connection Method - *eh+onecnct*
- Energy and Harmonicity Voicing Decision, Two-Connection Method - *eh+twocnct*
- Energy and Harmonicity Voicing Decision, Partial-Connection Method - *eh+partcnct*
- Energy and Harmonicity Voicing Decision, Partial-Connection Method, MFCC Distance Included - *eh+mfcc+partcnct*

As mentioned before, SUMMIT is somewhat sensitive to the placement of major landmarks, as small movements of the landmark locations can results in completely different word hypothesis and changes in error rates. Table 6.1 shows the error rates for the different segmentation methods on the AV-TIMIT development set contaminated by white noise.

In addition, the segmentation methods allow for different connections between segments, which results in different computation times. The processor computation

time for each segmentation method and SNR is shown in Figure 6-4. The recognizer shows greater confusability among speech models with increasing noise levels, resulting in less pruning and increased processor recognition time.

We will compare the different segmentation methods, both in terms of error rate and time, in the sections below.

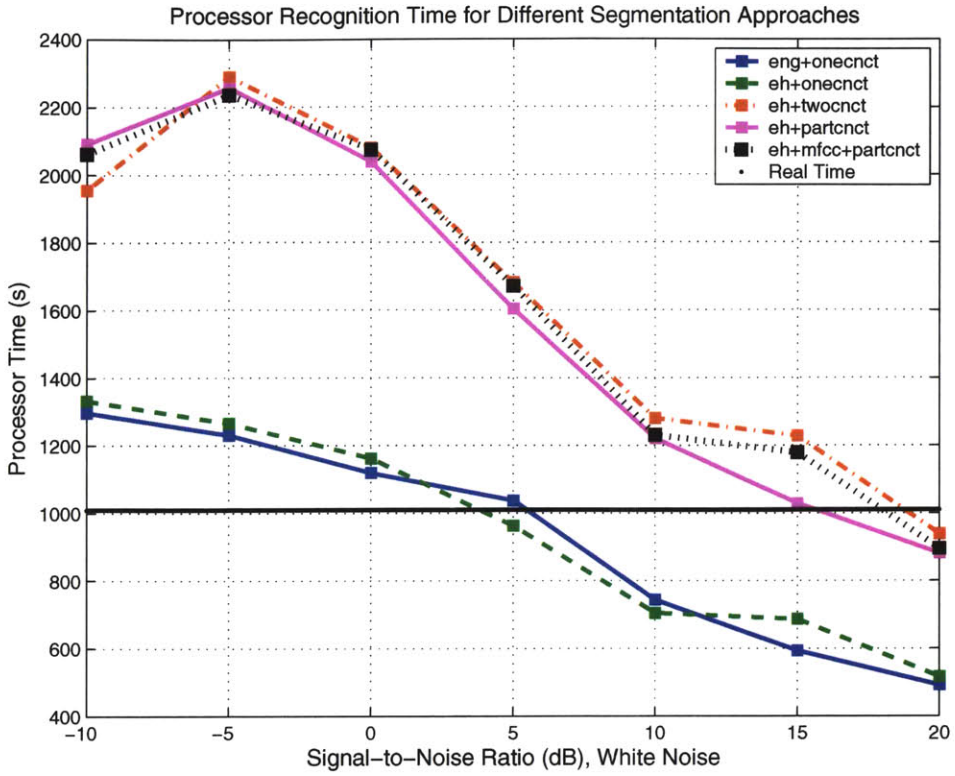

Figure 6-4: Recognition computation time for different segmentation approaches under varied SNR of white noise

6.3.1 Voicing Decision Methods

We explored hypothesizing major landmarks using short-time energy voicing decisions as well as using added harmonicity information. As can be seen from Table 6.1, hypothesizing major landmarks using energy and harmonicity as voicing decision improves the rate under each SNR. Short-time energy becomes less reliable and robust and robust in noisy environments as the signal weakens for particular phonemes,

Segmentation Approach	db level	wer	errors
eh+mfcc+partcnct	clean speech	1.3	31
eh+partcnct	clean speech	1.5	34
eh+twocnct	clean speech	1.3	31
eh+onecnct	clean speech	**1.2**	28
eng+onecnct	clean speech	1.4	33
eh+mfcc+partcnct	20db	**2.1**	49
eh+partcnct	20db	**2.1**	49
eh+twocnct	20db	**2.1**	49
eh+onecnct	20db	2.3	54
eng+onecnct	20db	**2.1**	49
eh+mfcc+partcnct	15db	**3.0**	70
eh+partcnct	15db	3.1	71
eh+twocnct	15db	3.2	74
eh+onecnct	15db	3.5	80
eng+onecnct	15db	4.1	95
eh+mfcc+partcnct	10db	**2.6**	59
eh+partcnct	10db	2.9	67
eh+twocnct	10db	3.1	71
eh+onecnct	10db	3.8	87
eng+onecnct	10db	4.3	98
eh+mfcc+partcnct	5db	**7.1**	164
eh+partcnct	5db	**7.1**	163
eh+twocnct	5db	7.4	170
eh+onecnct	5db	**7.1**	163
eng+onecnct	5db	7.9	181
eh+mfcc+partcnct	0db	14.1	326
eh+partcnct	0db	13.6	313
eh+twocnct	0db	**13.3**	306
eh+onecnct	0db	14.1	326
eng+onecnct	0db	17.7	409
eh+mfcc+partcnct	-5db	42.1	971
eh+partcnct	-5db	41.7	961
eh+twocnct	-5db	**41.0**	946
eh+onecnct	-5db	44.6	1028
eng+onecnct	-5db	42.6	981
eh+mfcc+partcnct	-10db	**96.8**	2232
eh+partcnct	-10db	96.9	2234
eh+twocnct	-10db	**96.8**	2232
eh+onecnct	-10db	97.1	2238
eng+onecnct	-10db	99.5	2294

Table 6.1: Word Error Rates for Different Sinusoidal Model Segmentation Approaches for varied SNRs of White Noise. Bold represents best segmentation method for each noise condition

making it difficult to accurately detect regions of voicing. Using harmonicity information helps to detect landmarks more precisely, thus improving word error rates. More importantly, the improved accuracy does not hurt the processor recognition time.

6.3.2 Connectivity Methods

Three different segment connectivity methods were also explored. Notice from Figure 6-4, the *eh+twocnct* and *eh+partcnct* methods have a much larger computation time over the *eh+onecnct* method, due to the greater segment connectivity. Since the energy and harmonicity combined measurement provided a lower word error rate than the short-time energy measurement alone, we compared the three connectivity methods using the combined measurements for the voicing decision.

At high SNRs, the *eh+twocnct* method improves the rate very little compared to a *eh+onecnct* connectivity connectivity approach. The signal energy is very strong and voicing decisions are more accurate at high SNRs. Therefore, increasing connectivity across major landmarks results in an increase in little improvement in word error rate. However, at low SNRs, the *eh+twocnct* method greatly improves the word accuracy in comparison to the *eh+onecnct* method.

The *eh+partcnct* method offers a nice tradeoff between error rate and computation time. As shown in Table 6.1, at high SNRs the word error rate is very similar to the the *eh+onecnct* and *eh+twocnct* methods. At low SNRs, word error rate is improved over the *eh+onecnct* method but not not as efficient as the *eh+twocnct* method. However, as shown in Figure 6-4, the computation time for the *eh+partcnct* method is lower compared to the *eh+twocnct* technique.

6.3.3 MFCC Distance

Finally, we investigated removing major landmarks placed in locations of little MFCC distance difference. Because the *eh+partcnct* technique offers the best tradeoff between accuracy and time, we apply this segment connectivity method to the landmarks generated from the MFCC distance method.

At high SNRs, the added MFCC distance information does not seem to help improve the word error rate compared to the previous methods. In fact, the *eh+mfcc+partcnct* method actually performs worse compared to the *eh+partcnct* technique. White noise has been shown to reduce the variance of cepstral coefficients [31]. Therefore, as the white noise level increases, the distance between MFCC feature vectors reduces. An increased number of major landmarks are removed which appear to have little change in spectral content, leading to a increased error rate. The removal of major landmarks increases segment connectivity and thus the computation time using the *eh+mfcc+partcnct* approach is also greater than the *eh+partcnct* method.

6.4 Comparison of Models

In this section, we will compare the performance, in terms of rate and computation time, of the sinusoidal model approach with the with the original and full segmentation techniques. Since the *eh+partcnct* segmentation method offers the best tradeoff between word error rate and computation time, we will apply this segmentation technique to the sinusoidal model approach in our experiments.

6.4.1 Word-Error-Rate

A certain landmark and segmentation method which is robust under one noise condition will not always perform well under different noise conditions. To test the noise robustness of the three methods, we compared the word error rates to white-noise, speech babble and destroyer operations room noise. Figures 6-5, 6-6 and 6-7 show the results under the three noise conditions. For more detailed description of these actual results, please refer to Appendix B.

With a finite amount of test data, it is often difficult to determine if the difference in performance of the three methods is statistically significant or not. Therefore, McNemar's significance test is often used test the statistical significance of the performance difference of two methods. In this thesis, if the statistical significance between two methods is less than 0.005, we say that the methods are statistically different.

Clean Speech			
	sinemodel	**origseg**	fullseg
sinemodel	= (1.000)	origseg	fullseg
origseg	origseg	= (1.000)	≈ (0.4373)
fullseg	fullseg	≈ (0.4373)	= (1.000)
10 dB			
	sinemodel	origseg	**fullseg**
sinemodel	= (1.000)	sinemodel	≈ (0.0072)
origseg	sinemodel	= (1.000)	fullseg
fullseg	≈ (0.0072)	fullseg	= (1.000)
0 dB			
	sinemodel	origseg	**fullseg**
sinemodel	= (1.000)	sinemodel	fullseg
origseg	sinemodel	= (1.000)	fullseg
fullseg	fullseg	fullseg	= (1.000)

Table 6.2: Comparison matrix showing results of McNemar's Test for Sinusoidal Model, Original Segmentation and Full Segmentation methods under varied SNRs of white noise. Methods which are statistically similar are indicated with an ≈ symbol and the corresponding significance level. If two methods are statistically different, the model with the better performance is indicated. Also, the model with the lowest error rate for each noise condition is indicated in bold.

The results for this test, applied to all three methods, are shown in Table 6.2.

The original, full and sinusoidal models all show similar behavior under all three noise conditions. At high SNRs, the full and original segmentation models tend to perform slightly better than the sinusoidal model. As shown in Table 6.2, the performance of the full and original segmentation methods under the clean speech condition are more statistically similar compared to the sinusoidal model.

At low SNRs, the performance of the original segmentation method rapidly degrades. However, the full segmentation method and sinusoidal model seem to degrade more gracefully. The performance of the original segmentation is significantly different compared to the full segmentation and sinusoidal model methods.

In addition, under all three noise conditions, the sinusoidal model provides a significantly better word error rate than the original segmentation method, but does not offer an improved error rate over the full segmentation model. However, the sinusoidal model is more statistically similar compared to the full segmentation method

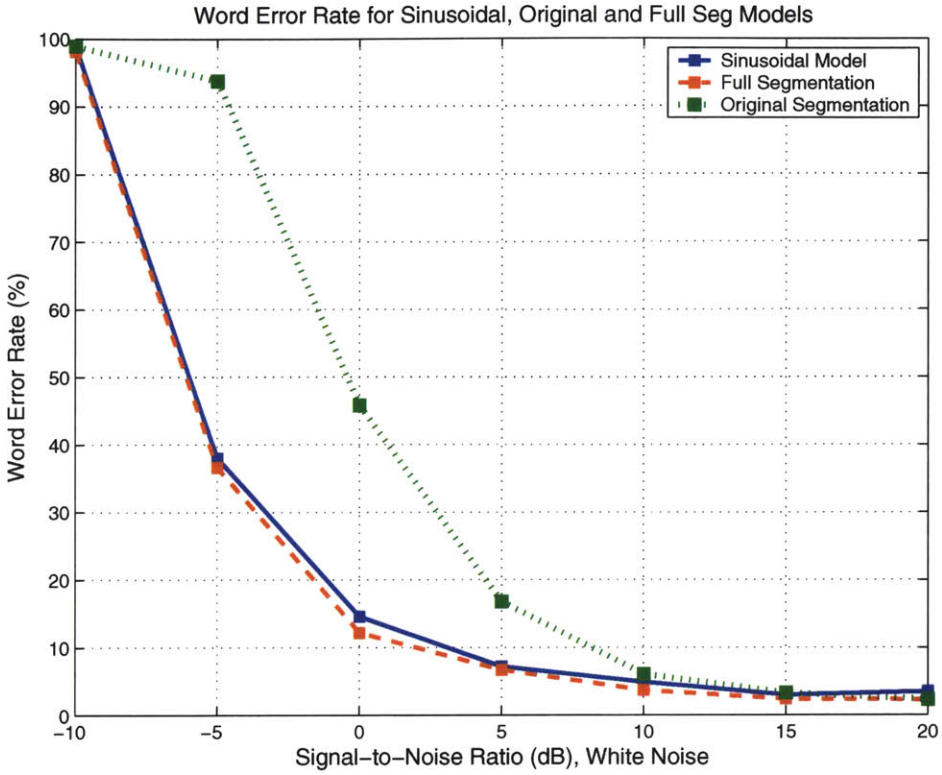

Figure 6-5: Word Error Rate for Original Segmentation, Full Segmentation and Sinusoidal Models for varied SNRs of White Noise

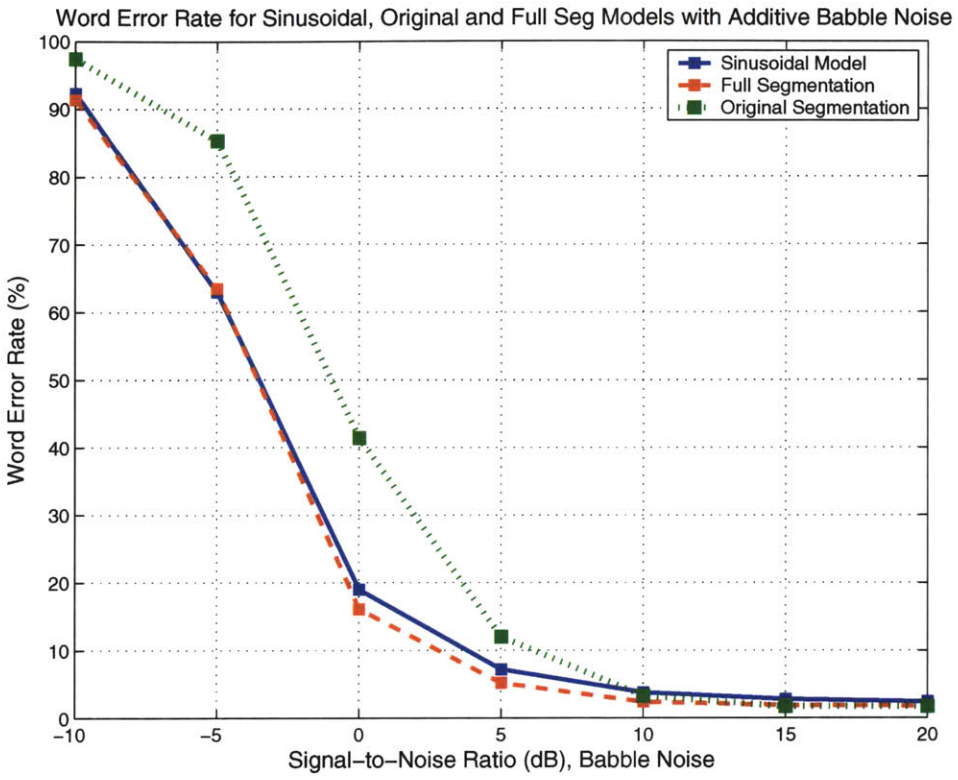

Figure 6-6: Word Error Rate for Original Segmentation, Full Segmentation and Sinusoidal Models for varied SNRs of Babble Noise

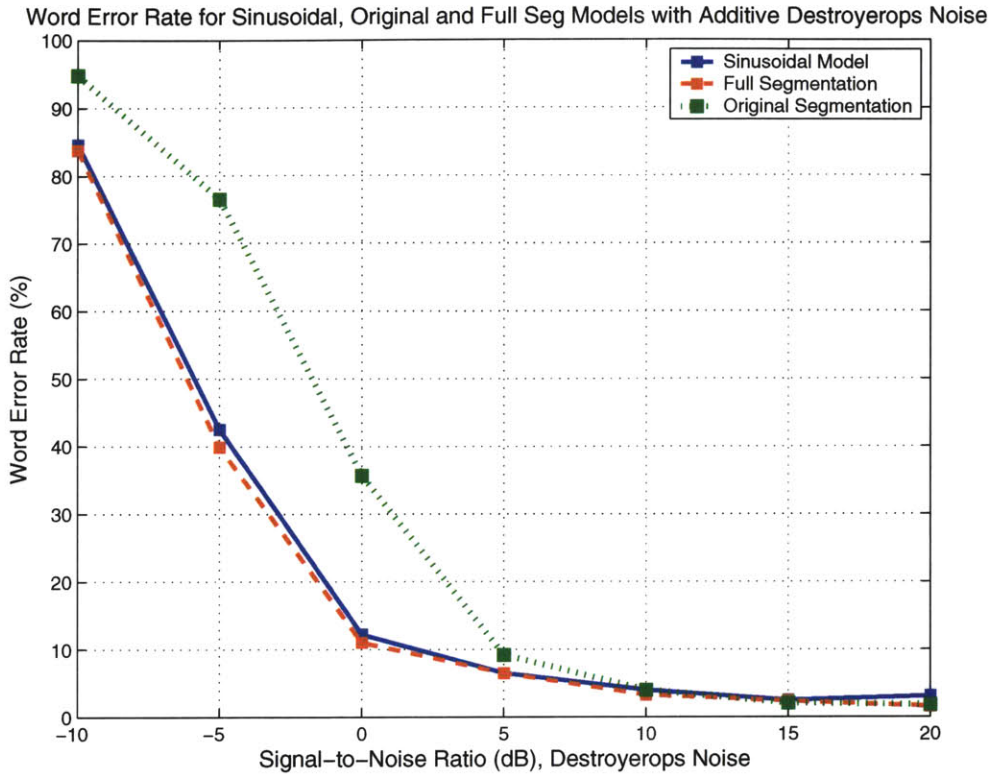

Figure 6-7: Word Error Rate for Original Segmentation, Full Segmentation and Sinusoidal Models for varied SNRs of Destroyerops Noise

than the original segmentation method, particularly under higher noise levels.

Finally, the sinusoidal approach appears to be robust to all three noise environments and does not rapidly degrade in any of the three environments. The model performs best when subject to destroyerops noise, since this noise type has a more sporadic and random characteristic than the other two noise types. White noise has a relatively flat spectrum and has similar sinusoidal characteristics as unvoiced speech. However, babble noise contains noise characteristics which are more similar to voiced speech than white noise. The babble noise will have a greater effect on the behavior of the sinusoidal components in voiced regions than white noise. Since we use these components to detect landmarks in voiced regions, this explains why the babble noise degrades the performance of the sinusoidal model more than white noise.

6.4.2 Recognition Computation Time

Each of the three techniques allow for different connections between segments, causing different recognition computation times. Figure 6-8 shows these times, in relation to real time, for the three methods as the SNR is varied under the white noise condition.

The full segmentation method provides little segmental constraint and therefore has the largest computation time. The original segmentation method computes major landmarks when the spectral change is above a specified global threshold. These landmarks are detected more often than the voicing decision landmarks of the sinusoidal model, explaining the lower computation time for the original segmentation method. However, the sinusoidal model segmentation allows for a much lower computation time compared to the full segmentation approach. The computation times for all three methods under the babble and destroyerops noise conditions also show similar trends.

The recognition computation time drops quickly for all three models at very high noise levels. The increased noise results in fewer landmarks detected for the original and sinusoidal models. In addition, the noise also causes the speech models to score poorly and possible word-sequence paths are pruned quicker for all three methods.

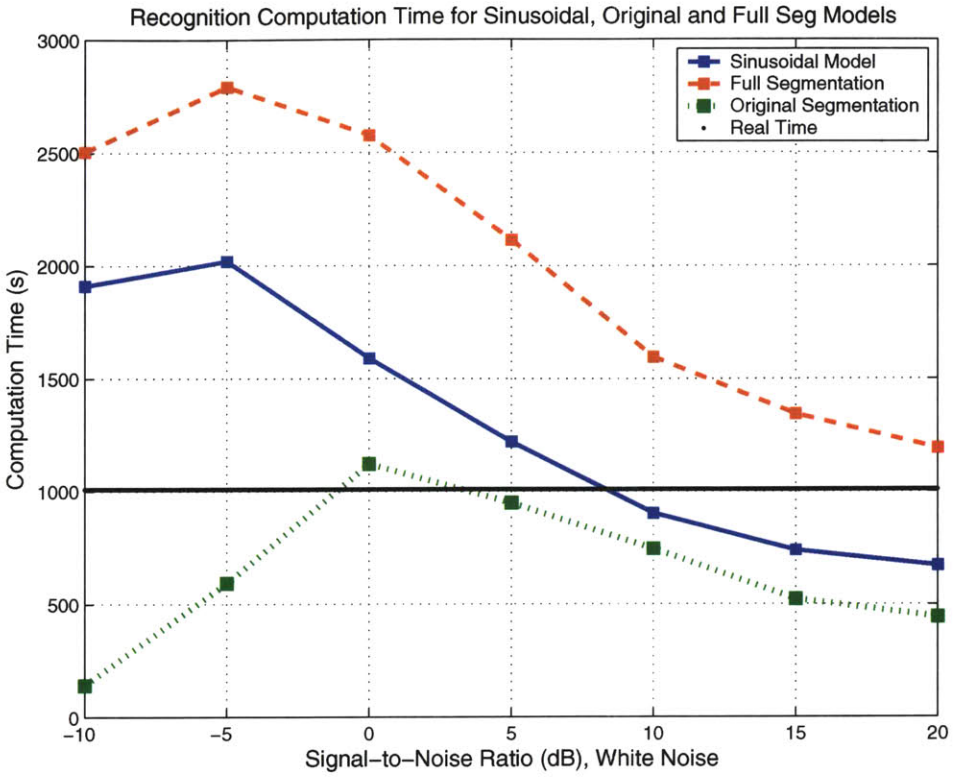

Figure 6-8: Recognition computation time for original, full and sinusoidal methods under varied SNR of white noise

6.4.3 Word Error Rate vs. Computation Time

Finally, to observe the tradeoff between word error rate and computation time for the three methods, we compute both statistics as we vary a Viterbi pruning threshold, known as *vprunenodes*. Pruning is done in the Viterbi search to remove unlikely paths from consideration and allow for a more efficient search time. At each step in the search, each possible word-sequence path is extended to the next state. A *vprunenodes* parameter is introduced to minimize the number of possible paths. At each stage, only the *vprunenodes* paths with the highest scores are extended and the rest are removed.

In our experiment, the three methods are compared at a SNR of 5dB, under each noise type. We vary *vprunenodes* from 100 to 10,000 and compute the word error rate and computation time for each *vprunenodes* setting. Figures 6-9, 6-10 and 6-11 illustrate the results for the three noise types in comparison to real time.

When the computation time is large, the sinusoidal and full segmentation models have a significantly lower word error rate compared to the original segmentation method. As the computation time is decreased, the full segmentation method has a faster increase in word error rate compared to the sinusoidal model. Finally, when the word error rate is high for all three methods, the sinusoidal model and original segmentation methods offer a much faster computation time than the full segmentation method. Thus, the sinusoidal model provides the best tradeoff between accuracy and time of the three methods under all three noise conditions.

The timing results in the figures above are plotted with respect to real time. For small tasks whose computation time is less than real time, the full segmentation method is best as it provides the lowest error rate of the three methods. However, for larger tasks which require more computation time, the error-timing curves in the above figures will move to the right as increased computation time is needed to achieve the same word error rate. While the full segmentation model will still have the lowest error, the computation time is greatly increased. The sinusoidal model has a slightly lower error rate than the full segmentation technique, but the smaller computation

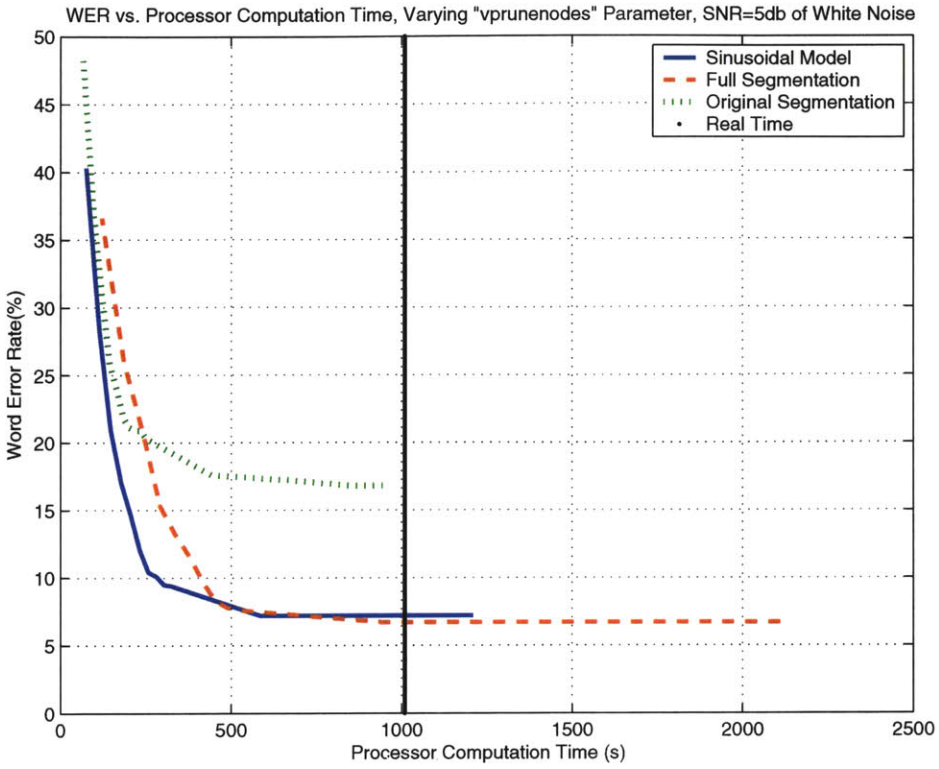

Figure 6-9: Word Error Rate vs. Computation Time for Original, Full and Sinusoidal Methods. Results are computed for varied *vprunenodes* parameter under 5dB of white noise.

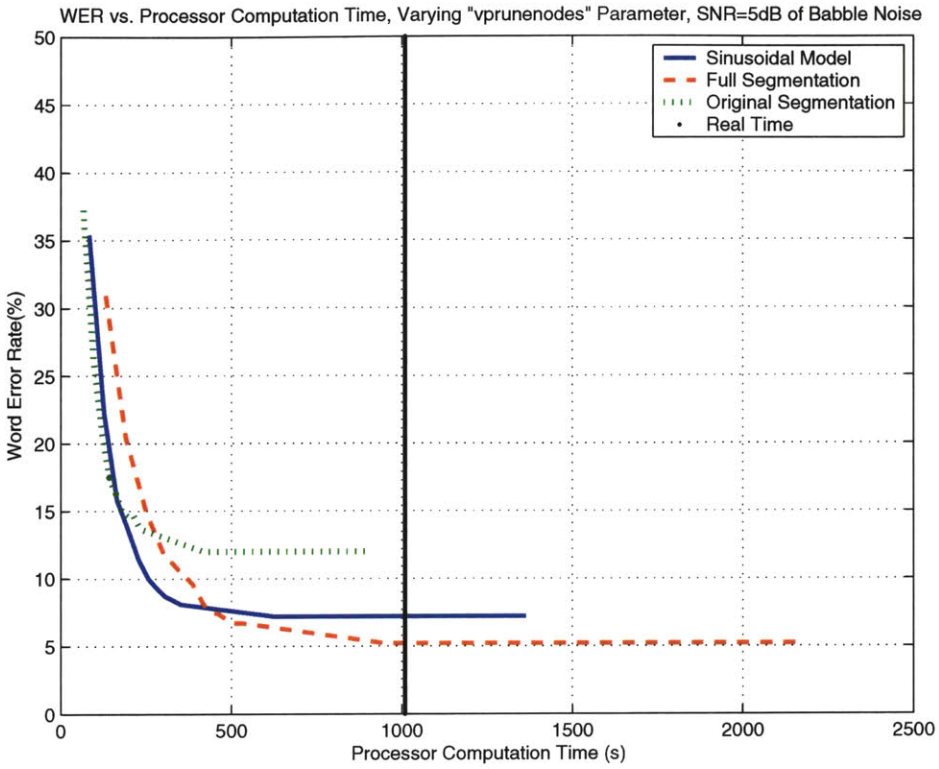

Figure 6-10: Word Error Rate vs. Computation Time for Original, Full and Sinusoidal Methods. Results are computed for varied *vprunenodes* parameter under 5dB of babble noise.

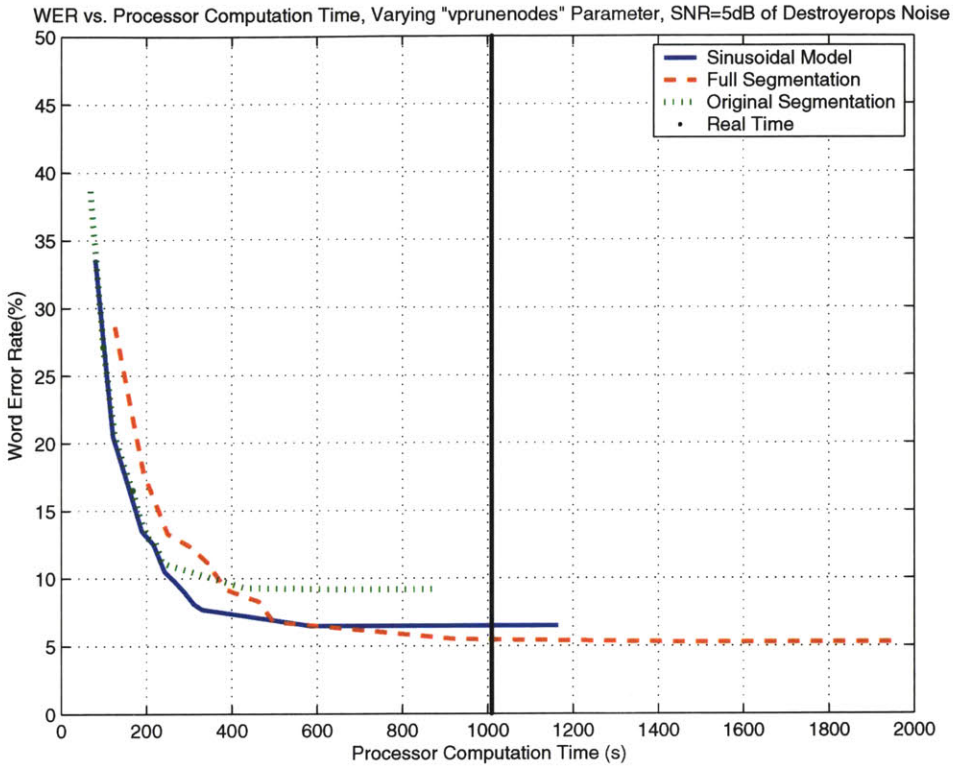

Figure 6-11: Word Error Rate vs. Computation Time for Original, Full and Sinusoidal Methods. Results are computed for varied *vprunenodes* parameter under 5dB of destroyerops noise.

time may make it more desirable for these larger tasks.

6.4.4 Landmark and Segment Comparisons

To gain a better understanding of the behavior of the original, full and sinusoidal methods, we analyzed the landmarks and segments generated from the three methods under different noise levels.

Landmark Accuracy

To observe the placement accuracy of hypothesized landmarks for each of the three methods, we compare the average time difference between landmarks hypothesized to the transcribed AV-TIMIT landmarks as estimated from forced alignments of clean AV-TIMIT utterances. For each of the AV-TIMIT landmarks, we find the closest hypothesized landmark and compute the time difference between these two landmarks. We then average the time difference across AV-TIMIT landmarks to compute an average landmark time difference. Figure 6-12 shows the landmark difference results for the three methods as we vary the SNR.

As the noise level increases, the original segmentation landmarks show a greater movement from the true AV-TIMIT landmarks compared to the full and sinusoidal model techniques. The original segmentation technique detects landmarks from changes in MFCC feature-vectors. The larger variance of MFCC feature vectors due to additive noise [31] is one explanation for the increased time difference of the original segmentation landmarks. Sinusoidal tracks still maintain their long, continuous behavior as the noise level is increased, resulting in a relatively constant average time landmark difference. The constant time difference of the full and sinusoidal models is one explanation for better performance over the original method at higher noise levels.

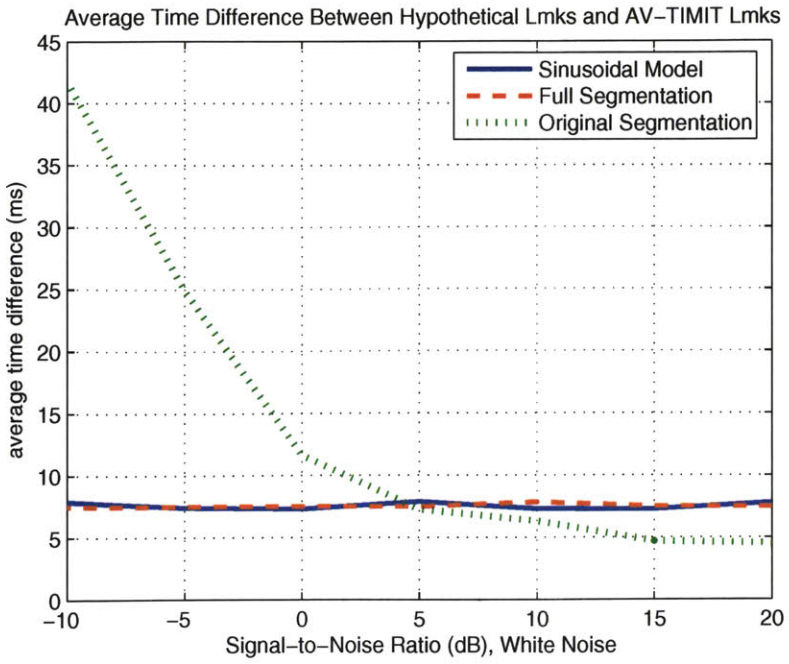

Figure 6-12: Average time difference between a hypothesized and AV-TIMIT landmark for original, full and sinusoidal methods under white noise

Segment Overgeneration

To observe the behavior of segments, we average the number of segments generated by the sinusoidal, original and full methods across each AV-TIMIT phone. Figure 6-13 illustrates the effect on segmentation for increased noise levels for the three methods.

At low noise levels, the original method overgenerates a smaller number of segments compared to the sinusoidal and full model. As explained in Section 6.4.2, the original method computes landmarks more frequently than the sinusoidal model, resulting in a smaller average number of segments per phone. The number of segments overgenerated by the original method shows a steady increase as the noise level is increased. Major segment boundaries (i.e. major landmarks) are computed at regions of large spectral change. Increased noise levels results in a smaller number of hypothesized major landmarks and a larger number of segments.

However, the number of segments for sinusoidal model changes at a much slower rate than the original model. The sinusoidal model method places major landmarks

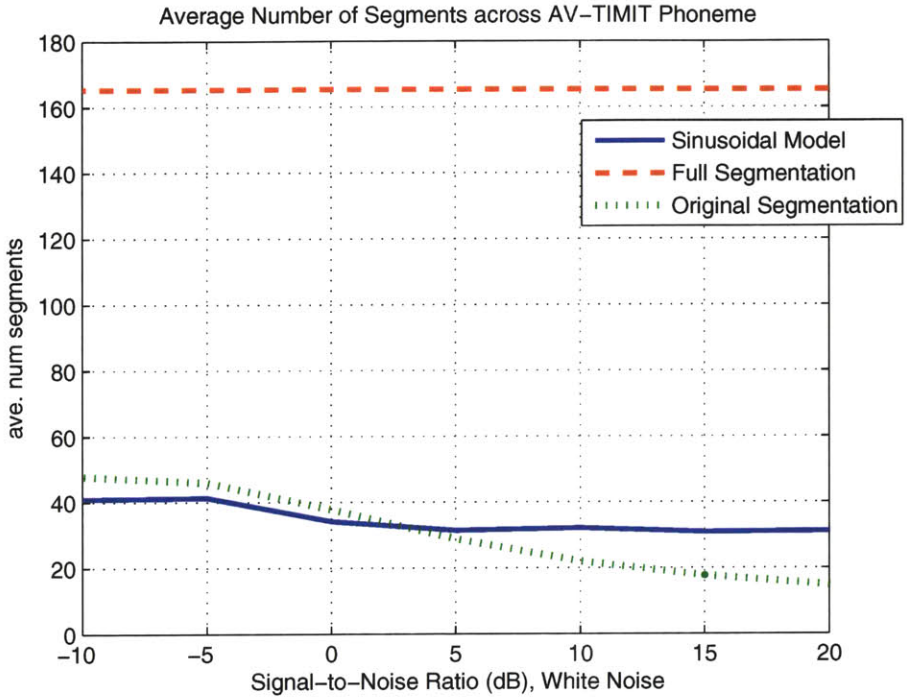

Figure 6-13: Average number of segments across an AV-TIMIT phoneme for original, full and sinusoidal methods under white noise

at voicing decision boundaries. As the noise level increases, it becomes harder to distinguish between voiced and unvoiced areas and accurately place major landmarks. However, the increased noise does not result in a large change in the number of major landmarks. Therefore, the sinusoidal model segmentation method is not as severely affected by increased noise as the original segmentation method.

Chapter 7

Conclusions

7.1 Summary

In this thesis, we explored a landmark detection and segmentation algorithm using the McAulay-Quatieri Sinusoidal Model. We compared the performance of our model to the original and full segmentation techniques in hopes of improving the word error rates and computation times, specifically under noisy speech environments.

7.1.1 Algorithm Design

Chapter 4 described our landmark detection method while the various segmentation methods explored are described in Chapter 5.

Landmark Detection

The McAulay-Quatieri Algorithm is used to represent our speech signal as a collection of sinusoidal components. The long, continuous behavior of sinusoids in voiced regions signaled phoneme transitions. However, the short, random behavior of the sinusoids in unvoiced regions provided little information about these transitions.

Short-time energy and harmonicity measurements were used to distinguish between voiced and unvoiced speech segments. In voiced regions, landmarks were detected from the births and deaths of these harmonic sinusoids. In unvoiced regions,

landmarks were placed at regularly spaced intervals. These hypothetical landmarks were interconnected together to form a collection of hypothetical segmentations of the utterance.

Segmentation Stage

After the segment graph is formed, our next step was to incorporate an explicit segmentation stage into this network to reduce the size of the search space. In this thesis, we explored many different segmentation methods.

As discussed in Section 5.4, major landmarks serve as hard boundaries for interconnections among minor landmarks. Major landmarks were hypothesized at voicing decision boundaries. In this thesis, we explored placing major landmarks when voicing decision boundaries are detected using a short-time energy measurement, as well as a combined short-time energy and harmonicity measurement. The combined measurement placed major landmarks more precisely and allowed for improved word error rates.

Secondly, we investigated various segment connectivity approaches, including a one-connection, two-connection and partial-connection method among landmarks. The two-partial connection method offered the best tradeoff between computation time and word error rates.

Finally, we explored removing major landmarks placed at regions of little MFCC distance. The MFCC distance information did not seem to significantly improve the word error rate and actually resulted in an increased computation time.

7.1.2 Performance of Sinusoidal Model

In Section 6.4, we compared the performance of our sinusoidal model, both in terms of word error rate and computation time, to the original and full segmentation techniques under noisy speech environments. Furthermore, to test the noise robustness of the sinusoidal model, we analyzed the performance under difference noise conditions.

Tradeoff of Word Error Rate and Computation Time

The sinusoidal model offered the best tradeoff between word error rate and computation time. At higher computation times, the sinusoidal and full segmentation models have a significantly lower word error rate than the original segmentation method. At high word error rates, the sinusoidal model and original segmentation methods offer a much faster computation time than the full segmentation method.

Noise Robustness

In addition, we analyzed the performance of the sinusoidal model when speech is contaminated by white noise, babble noise and destroyerops noise. The model appeared to be robust to all three noise environments and did not rapidly degrade under any of the three conditions.

7.2 Future Work

We would like to expand this work in a number of areas in the future. Humans are able to comprehend contaminated speech under many different noise conditions, yet few speech recognition systems can perform well in many different noise environments. In this thesis, we have taken the beginning steps towards achieving this noise robustness goal by demonstrating the robustness of the sinusoidal model under white nose, babble noise and destroyerops conditions. However, it would be interesting to study the performance of the sinusoidal model under other noise conditions. Since voiced sounds can be adequately estimated by a collection of sinusoids, we would particularly like to study the effect of adding periodic noise to the speech signal.

Furthermore, as shown in Section 6.4, the sinusoidal model performs poorly at very high noise levels. Section 1.2.1 discusses some speech enhancement techniques used to extract out clean speech from the contaminated utterance. It would be interesting to apply once of these techniques as a preprocessing step and then apply the sinusoidal model to the resulting clean speech.

In addition, the experiments in this thesis were performed on utterances from the AV-TIMIT corpus, with noise examples taken from the Noisex-92 corpus. We would like to further explore the performance of the sinusoidal model using another corpus. A popular database used in noise robust speech recognition experiments in the Aurora database [15]. This database includes suburban train noise, babble, car, exhibition hall, restaurant, street, airport and train-station noise

Finally, we would like to investigate using the sinusoidal model in an environment with many different sound types. Today, many speech recognition systems are able to successfully process human speech. However, little work has been done in adapting a system to detect and process non-speech sounds present in the environment, such as a horn beeping or a dog barking. Thus, people are limited to quiet environments where they can use these systems. Computational Auditory Scene Analysis (CASA) [2] is a process in which a system takes the mixture of sounds heard in a complex natural environment and sorts these sounds into packages of acoustic evidence in which each package corresponds to a sound class.

In the future, we would like to observe the performance of the sinusoidal model in an environment with many speech and non-speech sounds. The overlapping sounds from difference sources might correspond to different sinusoids, and we want to investigate if we can extract out these different sources by separating out individual sinusoids. If we can recognize these different sources, this will give us a more complete model of different acoustic sounds simultaneously present.

Appendix A

AV-TIMIT Phonemes

TIMIT	IPA	Example	TIMIT	IPA	Example
aa	ɑ	b*o*ttle	ix	ɨ	deb*i*t
ae	æ	b*a*t	iy	i	b*ee*t
ah	ʌ	b*u*t	jh	ʤ	*j*oke
ao	ɔ	b*ough*t	k	k	*k*ey
aw	ɑʷ	*ou*t	kcl	k˺	k closure
ax	ə	*a*bout	l	l	*l*ay
ax-h	əʰ	s*u*spect	m	m	*m*om
axr	ɚ	butt*er*	n	n	*n*oon
ay	ɑɪ	b*i*te	ng	ŋ	si*ng*
b	b	*b*ee	nx	ɾ̃	wi*n*ner
bcl	b˺	b closure	ow	o	b*oa*t
ch	č	*ch*oke	oy	ɔɪ	b*oy*
d	d	*d*ay	p	p	*p*ea
dcl	d˺	d closure	pau	▯	pause
dh	ð	*th*en	pcl	p˺	p closure
dx	ɾ	bu*tt*er	q	ʔ	co*tt*on
eh	ɛ	b*e*t	r	ɹ	*r*ay
el	l̩	bott*le*	s	s	*s*ea
em	m̩	bott*om*	sh	š	*sh*e
en	n̩	butt*on*	t	t	*t*ea
eng	ŋ̍	Washi*ng*ton	tcl	t˺	t closure
epi	▯	epenthetic silence	th	θ	*th*in
er	ɝ	b*ir*d	uh	ʊ	b*oo*k
ey	e	b*ai*t	uw	u	b*oo*t
f	f	*f*in	ux	ʉ	t*oo*t
g	g	*g*ay	v	v	*v*an
gcl	g˺	g closure	w	w	*w*ay
hh	h	*h*ay	y	y	*y*acht
hv	ɦ	a*h*ead	z	z	*z*one
ih	ɪ	b*i*t	zh	ž	a*z*ure
h#	-	utterance initial and final silence			

Figure A-1: 61 AV-TIMIT phones and corresponding International Phonetic Alphabet (IPA) Symbols, along with example words using the phonemes

Appendix B

Word Error Rate Tables

This Appendix lists the detailed word error rates on the AV-TIMIT test set for the sinusoidal, original and full segmentation methods, varying the SNRs of white, babble and destroyerops noise.

B.1 Word Error Rate for White Noise

Clean Speech		
method	word error rate	errors
sinemodel	3.1	72
fullseg	2.1	48
origseg	**1.8**	41
20db		
sinemodel	3.5	82
fullseg	**2.3**	54
origseg	**2.3**	54
15db		
sinemodel	3.0	71
fullseg	**2.4**	55
origseg	3.3	76
10db		
sinemodel	4.9	114
fullseg	**3.7**	86
origseg	6.1	156
5db		
sinemodel	7.2	169
fullseg	**6.7**	156
origseg	16.8	392
0db		
sinemodel	14.6	342
fullseg	**12.2**	285
origseg	45.8	1071
-5db		
sinemodel	38.0	888
fullseg	**36.6**	856
origseg	93.8	2191
-10db		
sinemodel	98.8	2307
fullseg	**98.1**	2292
origseg	98.9	2311

Table B.1: Word Error Rates for Original, Full and Sinusoidal Approaches for varied SNRs of White Noise. Bold represents best method for each noise condition.

B.2 Word Error Rate for Babble Noise

Clean Speech		
method	word error rate	errors
sinemodel	2.2	50
fullseg	1.7	39
origseg	**1.6**	37
20dB		
sinemodel	2.4	50
fullseg	1.8	43
origseg	**1.7**	40
15dB		
sinemodel	2.8	65
fullseg	1.9	45
origseg	**1.7**	40
10dB		
sinemodel	3.8	89
fullseg	**2.4**	56
origseg	3.3	76
5dB		
sinemodel	7.2	168
fullseg	**5.2**	121
origseg	12.0	280
0dB		
sinemodel	19.0	443
fullseg	**16.1**	377
origseg	41.4	966
-5dB		
sinemodel	**62.9**	1469
fullseg	63.4	1481
origseg	85.3	1992
-10dB		
sinemodel	92.3	2155
fullseg	**91.4**	2136
origseg	97.4	2276

Table B.2: Word Error Rates for Original, Full and Sinusoidal Approaches for varied SNRs of Babble Noise. Bold represents best method for each noise condition.

B.3 Word Error Rate for Destroyerops Noise

Clean Speech		
method	word error rate	errors
sinemodel	2.4	59
fullseg	1.5	36
origseg	**1.4**	33
20dB		
sinemodel	3.1	73
fullseg	**1.6**	38
origseg	1.8	43
15dB		
sinemodel	2.5	59
fullseg	2.4	59
origseg	**2.1**	49
10dB		
sinemodel	4.0	93
fullseg	**3.3**	77
origseg	4.0	93
5dB		
sinemodel	6.5	153
fullseg	**5.3**	123
origseg	9.2	214
0dB		
sinemodel	12.2	286
fullseg	**11.0**	257
origseg	35.7	834
-5dB		
sinemodel	42.5	992
fullseg	**39.9**	933
origseg	76.5	1787
-10dB		
sinemodel	84.6	1977
fullseg	**83.8**	1957
origseg	94.8	2214

Table B.3: Word Error Rates for Original, Full and Sinusoidal Approaches for varied SNRs of Destroyerops Noise. Bold represents the best method for each noise condition.

Bibliography

[1] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal. The AT&T Next-Gen TTS system. In *Proc. ASA*, Berlin, Germany, 1999.

[2] G.J. Brown and M.P. Cooke. Computational Auditory Scene Analysis. *Computer Speech and Language*, 8(4):297–336, 1994.

[3] M. J. Cheng, L. R. Rabiner, A. E. Rosenberg, and C. A. McGonegal. Some Comparisons Among Several Pitch Detection Algorithms. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 332–335, April 1976.

[4] D. P.W. Ellis. Lecture Notes from Speech and Audio Processing and Recognition Class, Lecture 6: Speech and Nonmusic. Columbia University Department of Electrical Engineering, Spr 2004.

[5] D.P.W Ellis. Sinewave and Sinusoid+Noise Analysis/Synthesis in MATLAB. http://www.ee.columbia.edu/ dpwe/resources/matlab/sinemodel/, 2003.

[6] K. Fitz, W. Walker, and L. Haken. Extending the McAulay-Quatieri Analysis for Synthesis with a Limited Number of Oscillators. *International Computer Music Conference*, 1992.

[7] M. Gales and S. Young. Robust Continuous Speech Recognition using Parallel Model Combination. In *IEEE Transactions on Speech and Audio Processing*, volume 4, September 1996.

[8] E.B. George. *An Analysis-by-Synthesis Approach to Sinusoidal Modeling Applied to Speech and Music Signal Processing.* PhD thesis, Georgia Institute of Technology, November 1991.

[9] J. Glass. *Finding Acoustic Regularities in Speech: Applications to Phonetic Recognition.* PhD thesis, Massachusetts Institute of Technology, 1988.

[10] J. Glass. A Probabilistic Framework for Segment-Based Speech Recognition. *Computer Speech and Language*, 17:137–152, 2003.

[11] J. Glass, T.J. Hazen, and L. Hetherington. Real-time Telephone-Based Speech Recognition in the JUPITER Domain. In *Proc. ICASSP*, pages 61–64, Phoenix, AZ, March 1999.

[12] Y. Gong. Speech Recognition in Noise Environments: A Survey. *Speech Communication*, 16(3):261–291, April 1995.

[13] T.J. Hazen, I.L. Hetherington, H. Shu, and K. Livescu. Pronunciation Modeling Using a Finite-State Transducer Representation. *Speech Communication*, 46(2):189–203, June 2005.

[14] T.J. Hazen, E. Saenko, C.H. La, and J. Glass. A Segment-Based Audio-Visual Speech Recognizer: Data Collection, Development and Initial Experiments. *Proc. of the International Conference on Multimodal Interfaces*, October 2004.

[15] H. G. Hirsch and D. Pearce. The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Condidions. In *ISCA ITRW ASR2000 "Automatic Speech Recognition: Challenges for the Next Millennium"*, Paris, France, September 2000.

[16] B.H. Juang. Speech Recognition in Adverse Environments. *Computer Speech and Language*, 5:275–294, 1991.

[17] J.C. Junqua, B. Mak, and B. Reaves. A Robust Algorithm for Word Boundary Detection in the Presence of Noise. *IEEE Transactions on Speech and Audio Processing*, 2(3):406–412, April 1994.

[18] S. Lee and J. Glass. Real-Time Probabilistic Segmentation for Segment-Based Speech Recogntion. In *Proc. ICSLP*, pages 1803–1806, Sydney, Australia, November 1998.

[19] R. J. McAulay and T. F. Quatieri. Speech Analysis/Synthesis Based on a Sinusoidal Representation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-34(4), August 1986.

[20] P.J. Moreno and R.M. Stern. Sources of Degradation of Speech Recognition in the Telephone Network. In *ICASSP*, pages 109–112, April 1994.

[21] A.V. Oppenheim and R.W. Schafer. *Discrete-Time Signal Processing.* Prentice-Hall, Inc., 2 edition, 1999.

[22] M. Ostendorf, V. Digalakis, and O. Kimball. From HMMs to Segment Models: A Unified View of Stochastic Modeling for Speech Recognition. *IEEE Transactions on Speech and Audio Proceesing*, 4(5):360–378, September 1996.

[23] T.F. Quatieri. *Discrete-Time Speech Signal Processing: Principles and Practice.* Prentice Hall, January 2002.

[24] L.R. Rabiner and M.R. Sambur. An Algorithm for Determining the Endpoints of Isolted Utterances. *The Bell System Technical Jounral*, 54(2):297–315, February 1975.

[25] X. Serra and J.O. Smith III. Spectral Modeling Synthesis: A Sound Analysis/Synthesis Based on a Deterministic plus Stochastic Decomposition. *Computer Music Journal*, 14, April 1990.

[26] K.N. Stevens. *Acoustic Phonetics.* The MIT Press, Cambridge, MA, 1998.

[27] Y. Stylianou, T. Dutoit, and J. Schroeter. Diphone Concatenation using a Harmonic plus Noise Model of Speech. In *Proc. EUROSPEECH*, pages 613–616, 1997.

[28] Y. Stylianou, J. Laroche, and E. Moulines. High-Quality Speech Modification based on a Harmonic + Noise Model. In *Proc. EUROSPEECH*, Madrid, Spain, 1995.

[29] H. Van Trees. *Detection, Estimation and Modulation Theory, Part I.* John Wiley and Sons, Hoboken, N.J, 2003.

[30] A.P. Varga, H.J.M. Steeneken, M. Tomlinson, and D. Jones. The NOISEX-92 Study on the Effect of Additive Noise on Automatic Speech Recognition. Technical report, Speech Research Unit, Defense Research Agency, Malvern, U.K., 1992.

[31] S. V. Vaseghi and B. P. Milner. Noise Compensation Methods for Hidden Markov Model Speech Recognition in Adverse Environments. *IEEE Transactions on Speech and Audio Processing*, 5(1):11–21, January 1997.

[32] V.W. Zue, S. Seneff, and J.R. Glass. Speech Database Development: TIMIT and Beyond. *Speech Communication*, 9(4):351–356, 1990.