# A SINUSOIDAL MODEL APPROACH TO ACOUSTIC LANDMARK DETECTION AND SEGMENTATION FOR ROBUST SEGMENT-BASED SPEECH RECOGNITION

*Tara N. Sainath and Timothy J. Hazen*

MIT Computer Science and Artificial Intelligence Laboratory
32 Vassar Street, Cambridge, MA 02139, USA
emails: {tsainath,hazen}@csail.mit.edu

## ABSTRACT

In this paper, we present a noise robust landmark detection and segmentation algorithm using a sinusoidal model representation of speech. We compare the performance of our approach under noisy conditions against two segmentation methods used in the SUMMIT segment-based speech recognizer, a full segmentation approach and an approach that detects segment boundaries based on spectral change. The word error rate of the spectral change segmentation method degrades rapidly in the presence of noise, while the sinusoidal and full segmentation models degrade more gracefully. However, the full segmentation method requires the largest computation time of the three approaches. We find that our new 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 various noise types.

## 1. INTRODUCTION

Hidden Markov Models (HMMs) have been the most dominant frame-based acoustic modeling technique for automatic speech recognition tasks to date. However, alternative models, such as segment-based models, have been developed to address the limitations of HMMs [1]. For example, the SUMMIT speech recognizer uses a segment-based framework for acoustic modeling [2]. This system computes a temporal sequence of frame-based feature vectors from the speech signal, and performs spectral energy change based landmark detection. These landmarks, representing possible transitions between phones, 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 existing algorithm works well in clean conditions as well as telephony applications [3].

In recent years, improvements in speech recognition systems have resulted in high performance on specific tasks under clean conditions. However, the performance of these systems can rapidly degrade in noisy environments [4]. Similarly, the spectral change segmentation algorithm used in SUMMIT performs poorly in the presence of strong background noises and non-speech sounds. Specifically, the system has difficulty locating landmarks in the presence of noise and often produces poor segmentation hypotheses.

We have observed that noise robustness can be improved using a full segmentation method (i.e., an exhaustive segmental search). 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 spectral segmentation method, it is more robust under noisy environments.

To address the limitations of the spectral change and full segmentation methods, we have developed a new landmark detection and segmentation algorithm from the behavior of sinusoidal components generated from the McAulay-Quatieri Sinusoidal Model [5]. Our goal is the development of a robust method which provides a good tradeoff between word error rate and computation time under different noise environments. Specifically, we hope to improve upon the word error rate of the spectral segmentation method while providing faster computation time than the full segmentation method.

In the following section, we describe our landmark detection and segmentation algorithm. Section 3 presents the experiments performed, followed by a discussion of these results in Section 4. Finally, Section 5 concludes the paper and discusses future work.

## 2. LANDMARK DETECTION AND SEGMENTATION

### 2.1. McAulay-Quatieri Sinusoidal Model

The McAulay-Quatieri (MQ) Algorithm is often used to produce a sinusoidal representation of sounds [5]. For this work, we use a MATLAB implementation of the MQ Sinusoidal Model developed by Ellis.[1] The algorithm assumes that a speech waveform can be represented by a collection of sinusoidal components of arbitrary amplitudes, frequencies and phases. 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. The speech signal analyzed with the sinusoidal model, can be expressed as follows:

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

When examining speech waveforms, it is observable that voiced sounds can adequately be estimated by a harmonic collection of sinusoids whose peaks in the Short-Time Fourier Transform occur at close amplitudes and frequencies from frame to frame. Because sinusoidal tracks are connected by matching peaks at close frequencies in contiguous frames, the proximity in peak frequencies between frames results in long-duration tracks with peaks whose amplitude and frequency variations are smooth and slowly varying. The births and deaths of these long, continuous sinusoids in voiced regions typically occur at phoneme transitions.

According to the Karhunen-Loève analysis, unvoiced signals can only be sufficiently modeled by a very large number of sinu-

This work was sponsored in part by an industrial consortium supporting the MIT Oxygen Alliance and in part by the Quanta-MIT T-Party Alliance.

[1] http://www.ee.columbia.edu/ dpwe/resources/matlab/sinemodel/

soids. 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, rapidly fluctuating tracks. The sinusoidal births and deaths often occur too frequently and randomly to signal a phonetic transition.

### 2.2. Landmark Detection

In this section, we describe our method for detecting landmarks from sinusoidal components. Because only sinusoids in voiced regions are useful for landmark detection, we first describe our method to distinguish voiced and unvoiced regions of the speech utterance, and then proceed to discuss our method of landmark detection in each of these regions.

#### *2.2.1. Voicing Detection Method*

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

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

where $n$ is the center sample of the windowed speech region, $s[m]$ are the speech samples, and $w[n-m]$ corresponds to the window. In voiced regions the signal energy is typically high, while in unvoiced regions the signal energy is usually low. While the short-time energy has been conventionally used to distinguished between voiced and unvoiced segments, this measure becomes less reliable in noisy environments. Specifically, in noisy environments it becomes difficult to accurately detect regions of voicing with a low signal-to-noise ratio.

Harmonicity, a measure of the strength of the pitch perception for a sound, can also be used to distinguish between voiced and unvoiced speech regions. Harmonicity can be calculated as the ratio of harmonic energy, $e_h[n]$, to the total signal energy, as follows:

$$h[n] = \frac{e_h[n]^2}{e[n]^2}, \quad 0 < h[n] < 1 \tag{3}$$

where $e_h[n]$ is determined from the harmonic components of the automatically estimated fundamental frequency of the speech signal.

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. In regions of weak voicing where the short-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 regions of voicing, but the harmonicity can help to make the locations of these regions more precise. In our approach, sharp changes in harmonicity in conjunction with changes in short-time energy are used to hypothesize voicing change landmarks.

#### *2.2.2. Landmark Detection Method*

In order to determine the best method for detecting phonetic landmarks, we look at the behavior of sinusoids with respect to the actual phonetic boundaries obtained from forced transcriptions of clean speech. In voiced regions, we hypothesize phonetic landmarks by examining a set of features found useful for landmark detection. The most useful features are the number of harmonically related sinusoids that are born or die within a set frame interval (see [6] for details). We also set various parameters for the minimum number of frames required between hypothesized landmarks. Landmarks are hypothesized when specified features, such as the number of new sinusoids born at a given frame, exceed a predetermined threshold. The thresholds are determined by examining the receiver-operating characteristic (ROC) of landmark detection and finding suitable settings which offer a high landmark detection rate while limiting the landmark over-generation rate.

As stated previously, in unvoiced regions sinusoidal births and deaths occur too frequently to indicate phonetic transitions. Furthermore, the full segmentation approach, which places potential landmarks at a fixed interval, has a much lower word error rate in the presence of noise than the spectral segmentation approach. Therefore, in unvoiced regions we decided to place landmarks at a fixed frame interval.

### 2.3. Segmentation

After landmarks are detected, they are interconnected together to form a network of hypothetical segmentations. It is computationally expensive to search a segmentation network that fully connects all hypothesized landmarks. Thus an explicit segmentation phase is used to reduce the search space by removing segments that are unlikely to correspond to single phonetic units. While the segmentation phase reduces the computation time of the recognizer, excessive pruning of the segment network can result in the deletion of actual phonetic segments. Thus, we have explored a variety of different segmentation methods [6].

To minimize landmark interconnections, we label the subset of voicing landmarks as *major* landmarks. These voicing landmarks are placed at locations predicted to contain a change in voicing. In [6], we detail several methods for predicting voicing landmarks based on short-time energy, harmonicity, and spectral change across the potential landmarks. In this paper, we utilize a metric that combines short-time energy and harmonicity. Furthermore, in this paper we reclassify these major voicing landmarks as *hard major* or *soft major* landmarks based on the energy difference across the landmark. All other predicted landmarks within voiced and unvoiced regions are termed *minor* landmarks.

In [6] we also detail various methods for connecting landmarks to generate a segmentation graph. In this paper, our experiments use a connectivity algorithm with the following rules:

1. Segments starting at a minor landmark can end at any future landmark as long as no hard major landmarks are crossed and at most one soft major landmark is crossed.
2. Segments starting at a major landmark can end at any future minor landmark as long as no hard major landmarks are crossed and at most one soft major landmark is crossed.
3. Segments starting at a major landmark can end at any future major landmark as long as at most one major landmark is crossed.

The spectral change segmentation algorithm used by SUMMIT uses similar connectivity rules to those expressed above, though it makes no distinction between soft and hard major boundaries. Our full segmentation approach places landmarks at fixed intervals of 30ms apart with full connectivity allowed up to 250ms away. Experiments using full segmentation at 10ms intervals was explored in [6], but performed considerably worse than the 30ms interval case in both accuracy and computation time.

## 3. EXPERIMENTS

### 3.1. Corpora

Our recognition experiments draw from two distinct corpora. The AV-TIMIT corpus is a collection of speech recordings developed for research in audio-visual speech recognition [7]. The corpus consists of phonetically balanced utterances based on SX sentences drawn from the TIMIT Corpus. The vocabulary for this corpus consists of 1793 words. In addition, the language model uses 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, resulting in an average perplexity of 3.

We simulate noisy speech by adding noise from the Noisex-92 database [8] to clean AV-TIMIT utterances at signal-to-noise ratios in the range of -10db to 20db in 5db increments. In this work we look at three specific types of noises, white-noise, speech babble and destroyer operations room noise.

We also performed experiments using the AURORA 2 database [9], which consists of clean TI-digit utterances with artificially added noise at levels of -5db to 20db in 5db increments. For this work, we report results only on Test Set A, which contains noise types similar to those seen in the training data, namely subway, babble, car, and exhibition hall noise.

### 3.2. Experimental Setup

Our experiments compare word error rate and recognizer computation times using the SUMMIT recognizer with the sinusoidal model, full, and spectral segmentation methods for both corpora. Furthermore, to observe the tradeoff between word error rate and computation time for the three methods, we compute both statistics as we vary the Viterbi pruning threshold which limits the number of possible paths at each step in the recognition search. For AV-TIMIT, each experiment uses acoustic models matched to the test data's SNR and noise type. For Aurora 2, global multistyle acoustic modeling is used for all experiments.

## 4. RESULTS

### 4.1. Word Error Rate

Tables 1 and 2 show that, as the noise level increases, the performance of the spectral segmentation method degrades rapidly compared to the full and sinusoidal model segmentation methods. The spectral segmentation technique, which detects landmarks from differences between adjacent MFCC feature-vectors, becomes ineffective in regions of small signal-to-noise ratios. However, the sinusoidal method is able to detect long, continuous tracks of harmonically related sinusoids even as the noise level is increased. Discussion on this observation can be found in [6]. The full segmentation method generally performs better than the sinusoidal model for the AV-TIMIT task, but the converse is true for the Aurora task.

### 4.2. Word Error Rate vs. Computation Time

Figures 1 and 2 illustrate the tradeoff between word error rate and computation time under noisy conditions as the pruning threshold is varied for the AV-TIMIT and Aurora tasks. When the computation time is large, the sinusoidal and full segmentation approaches have a significantly lower word error rate than the spectral segmentation method. As the computation time is decreased, the word error rate of the full segmentation method increases sooner than the sinusoidal model approach. Finally, when the word error rate is high for all three methods, the sinusoidal model and spectral segmentation methods offer a much faster computation time than the full segmentation method. Thus, the sinusoidal model provides the best tradeoff between accuracy and computation time under all noise conditions.

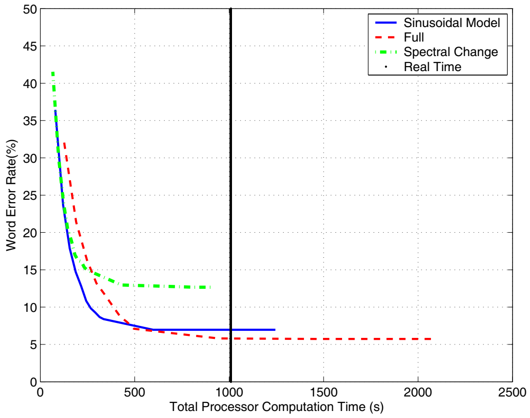

**Fig. 1**. Word error rate vs. computation time for spectral, full and sinusoidal methods on the AV-TIMIT corpus averaged over the 3 noise conditions at a signal-to-noise ratio of 5dB.

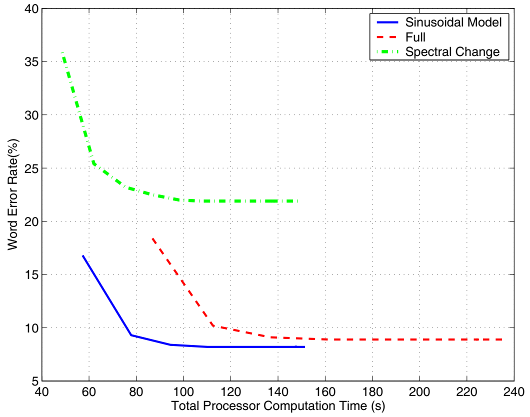

**Fig. 2**. Word error rate vs. computation time for spectral, full and sinusoidal methods on the Aurora 2 corpus averaged over the 4 noise conditions at a signal-to-noise ratio of 10dB.

### 4.3. Noise Robustness

Finally, the sinusoidal approach appears to be robust and does not rapidly degrade under any of the noise environments. The sinusoidal model performs best when subject to sporadic noise conditions (e.g., destroyer operations). White, car and subway noise have a relatively flat spectrum with sinusoidal characteristics similar to unvoiced speech. However, the babble and exhibition noise conditions have characteristics which are similar to voiced speech. These noises have a greater effect on the behavior of the sinusoidal components in voiced regions than white and motor noises, which might explain the decrease in performance of the sinusoidal model compared to the unvoiced noise types.

| AV-TIMIT Test Results | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | White | | | Babble | | | Destroyer Operations | | | Average | | |
| dblevel | sine | full | spec | sine | full | spec | sine | full | spec | sine | full | spec |
| Clean | 1.5 | 1.1 | 1.1 | 2.2 | 1.7 | 1.1 | 2.4 | 1.4 | 1.5 | 2.0 | 1.4 | 1.2 |
| 20db | 2.1 | 1.8 | 2.0 | 2.4 | 1.8 | 1.7 | 3.1 | 1.6 | 1.8 | 2.5 | 1.7 | 1.8 |
| 15db | 3.1 | 2.5 | 3.0 | 2.8 | 1.9 | 1.7 | 2.5 | 2.4 | 2.1 | 2.8 | 2.3 | 2.3 |
| 10db | 2.9 | 3.3 | 7.3 | 3.8 | 2.4 | 3.3 | 4.0 | 3.3 | 4.0 | 3.6 | 3.0 | 4.9 |
| 5db | 7.1 | 6.6 | 19.8 | 7.2 | 5.2 | 12.0 | 6.5 | 6.5 | 9.2 | 6.9 | 6.1 | 13.7 |
| 0db | 13.6 | 13.4 | 54.6 | 19.0 | 16.1 | 41.1 | 12.2 | 11.0 | 35.7 | 14.9 | 13.5 | 43.8 |
| -5db | 41.7 | 37.2 | 95.2 | 62.9 | 63.4 | 85.3 | 42.5 | 29.9 | 76.5 | 49.0 | 43.5 | 85.7 |
| -10db | 96.9 | 98.1 | 98.7 | 92.3 | 91.4 | 97.4 | 84.6 | 83.8 | 94.8 | 91.3 | 91.1 | 97.0 |
| Average | 21.1 | 20.5 | 35.2 | 24.1 | 23.0 | 30.5 | 19.7 | 17.5 | 28.2 | 21.6 | 20.3 | 31.3 |

**Table 1**. Word error rates for spectral, full and sinusoidal methods using the AV-TIMIT corpus.

| Aurora 2 Test Results | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Subway | | | Babble | | | Car | | | Exhibition | | | Average | | |
| db | sine | full | spec | sine | full | spec | sine | full | spec | sine | full | spec | sine | full | spec |
| Clean | 3.1 | 2.0 | 2.4 | 2.8 | 2.0 | 2.3 | 3.4 | 1.9 | 2.4 | 3.0 | 1.4 | 1.9 | 3.1 | 1.8 | 2.3 |
| 20db | 2.5 | 2.1 | 3.0 | 3.5 | 3.4 | 4.5 | 2.8 | 2.6 | 3.0 | 4.2 | 3.5 | 4.7 | 3.3 | 2.9 | 3.8 |
| 15db | 2.6 | 2.2 | 5.8 | 5.1 | 5.0 | 8.2 | 3.8 | 3.0 | 4.0 | 6.0 | 6.1 | 10.8 | 4.4 | 4.1 | 7.2 |
| 10db | 5.1 | 5.7 | 16.0 | 9.2 | 10.2 | 22.5 | 6.4 | 6.5 | 21.1 | 12.2 | 13.4 | 26.8 | 8.2 | 9.0 | 21.6 |
| 5db | 13.9 | 15.9 | 39.6 | 25.8 | 31.3 | 55.5 | 16.1 | 17.7 | 54.4 | 26.2 | 28.1 | 59.6 | 20.5 | 23.3 | 52.3 |
| 0db | 35.6 | 39.7 | 72.0 | 60.6 | 71.5 | 88.5 | 41.8 | 48.2 | 81.8 | 54.4 | 59.5 | 84.4 | 48.1 | 54.7 | 81.7 |
| -5db | 66.7 | 75.4 | 86.7 | 87.3 | 95.9 | 97.8 | 75.3 | 82.0 | 90.5 | 81.8 | 85.8 | 93.8 | 77.8 | 84.8 | 92.2 |
| Average | 18.5 | 20.4 | 32.2 | 27.8 | 31.3 | 39.9 | 21.4 | 23.1 | 36.7 | 26.8 | 28.3 | 40.3 | 23.6 | 25.8 | 37.3 |

**Table 2**. Word error rates for spectral, full and sinusoidal methods using the Aurora 2 corpus.

## 5. CONCLUSIONS AND FUTURE WORK

In this work, we explored a landmark detection and segmentation algorithm using a sinusoidal model. We found that our method offered the best tradeoff between word error rate and recognition computation compared to the spectral and full segmentation methods when used within the SUMMIT segment-based recognition system. Furthermore, our method was robust to the various different noise environments used in our experiments.

We would like to expand this work in a number of areas in the future. Since voiced sounds can be adequately estimated by a collection of sinusoids, we would like to study the effect of adding periodic noise to the speech signal. Furthermore, we would like to observe the performance of the sinusoidal model in realistic environments which may contain a variety of background speech and non-speech sounds. First, we would like to examine the behavior of sinusoidal tracks under different noise conditions to see if we can classify acoustic regions into different sound classes. Secondly, overlapping sounds from difference sources might correspond to different sinusoidal tracks. We would like to explore the use of source separation techniques based on the sinusoidal model as a potential means of improving the robustness of our acoustic models via speech enhancement.

## 6. REFERENCES

[1] 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*, vol. 4, no. 5, pp. 360–378, September 1996.

[2] J. Glass, “A Probabilistic Framework for Segment-Based Speech Recognition,” *Computer Speech and Language*, vol. 17, pp. 137–152, 2003.

[3] J. Glass, T.J. Hazen, and I.L. Hetherington, “Real-Time Telephone-Based Speech Recognition in the Jupiter Domain,” in *Proc. ICASSP*, Phoenix, AZ, March 1999.

[4] Y. Gong, “Speech Recognition in Noise Environments: A Survey,” *Speech Communication*, vol. 16, no. 3, pp. 261–291, April 1995.

[5] R. J. McAulay and T. F. Quatieri, “Speech Analysis/Synthesis Based on a Sinusoidal Representation,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-34, no. 4, August 1986.

[6] T.N. Sainath, “Acoustic Landmark Detection and Segmentation Using the McAulay-Quateiri Sinusoidal Model,” M.S. thesis, Massachusetts Institute of Technology, August 2005.

[7] 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.

[8] 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,” Tech. Rep., Speech Research Unit, Defense Research Agency, Malvern, U.K., 1992.

[9] 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.