# AUTOMATIC SEGMENTATION AND LABELING OF SPEECH RECORDED IN UNKNOWN NOISY CHANNEL ENVIRONMENTS *

*Bryan L. Pellom* and *John H.L. Hansen*

Robust Speech Processing Laboratory
Department of Electrical and Computer Engineering
Duke University, Box 90291, Durham, NC 27708-0291, USA
http://www.ee.duke.edu/Research/Speech bp@ee.duke.edu jhlh@ee.duke.edu

## ABSTRACT

This paper investigates the problem of automatic segmentation of speech recorded in noisy channel corrupted environments. Several speech enhancement and channel compensation techniques previously proposed for robust speech recognition are evaluated and compared for improved segmentation in colored noise. Speech enhancement algorithms considered include: Generalized Spectral Subtraction, Nonlinear Spectral Subtraction, Ephraim-Malah MMSE enhancement, and AutoLSP Constrained Iterative Wiener filtering. In addition, the coupling of front-end processing such as cepstral mean subtraction in tandem with speech enhancement methods is considered for noisy channel compensation. Compensation performance is assessed for each method using automatic segmentation of the telephone transmitted NTIMIT and cellular telephone transmitted CTIMIT databases.

## 1. INTRODUCTION

Segmentation and labeling of speech is both a time-consuming and tedious problem. Often, the exact position of phoneme boundaries are disputed even by expert phoneticians. As a result, automated procedures for segmenting speech have been proposed in order to improve the efficiency and accuracy of speech database collection [1, 2]. In recent years there has been wide interest in investigating methods for environmental and channel robustness for improved speech recognition. Such progress has produced the need for new development and validation databases collected in telephone, cellular, or noisy environments. Unfortunately, many segmentation systems have been formulated to align speech recorded in ideal clean conditions, and consequently, have reduced performance in a mismatched environment. In this paper, we present an HMM-based speech time-alignment system. The initial segmentation system is then used to investigate the impact of speech enhancement and channel compensation methods for improving automatic labeling accuracy in a variety of non-ideal recording environments. Specifically, we consider additive noise conditions such as aircraft cockpit and automobile highway noise in addition to telephone recording environments. For the latter, segmentation accuracy is examined using the NTIMIT and CTIMIT databases for a combination of front-end processing schemes including cepstral mean subtraction, speech enhancement, and model training using a wireline channel simulator.

*This research was supported in part by a National Science Foundation Graduate Research Fellowship.

## 2. SPEECH SEGMENTATION ALGORITHM

The speech segmentation system used in this study is based on continuous density hidden Markov models (HMMs). In the baseline system, 46 phoneme units are modeled using 5-state left-to-right HMMs. For each state, 16 mixture densities are used to characterize the observation probability density. During segmentation, the speech waveform is parameterized by a vector consisting of 12 mel-cepstral, 12 delta mel-cepstral, and normalized log-frame energy. The parameter sequence is extracted every 5 msec around a 25 msec frame window. Given a sequence of phonemes to align, a composite HMM is constructed by concatenation of unit phoneme models. The state sequence through the composite model is then determined using the Viterbi algorithm. Phoneme boundaries are noted for each transition from the proceeding unit HMM to that of the succeeding unit HMM along the maximum likelihood state path.

In addition to phoneme-based segmentation, we have integrated the *CMU Pronouncing Dictionary* with the alignment tool in order to obtain phoneme transcriptions from word-level transcriptions. This dictionary contains phonetic transcriptions of over 115,000 common American English words. In the case of word-based alignment, the phonetic spelling of each word is determined using a dictionary lookup. Additionally, we have implemented a word-juncture modeling [3] routine in order to obtain more accurate phonetic sequences from word-level transcriptions. For example the phrase "what time" could be transcribed as /w ah t ay m/ rather than /w ah t t ay m/. The resulting phoneme sequence is then used to construct the composite HMM. When multiple word pronunciations are encountered, all possible phonetic sequences are scored in order to determine the most probable state sequence (i.e., time-alignment) and pronunciation for the input utterance.

In developing this system, we experimented with using 3 and 5-state HMMs in addition to gender dependent HMMs. It was found that a 5-state HMM configuration outperformed that of 3-state models and that gender dependent HMMs further improved segmentation accuracy over gender-independent models. The model parameters were estimated using speech data from the training portion of an 8 kHz sampled version of the TIMIT database. The algorithm was tested by forced alignment of speech contained in the core test set of the TIMIT database. The results of the automatic segmentation for this ideal clean environment are given in Table 1. Here we see that 80.3 % of the automatically segmented boundaries were placed within 15 msec of the actual hand-labeled location. These results are comparable to those reported using Entropic's *Aligner* on the 16 kHz sampled TIMIT database (71% < 16 msec, 90% < 32 msec, 97% < 64 msec). We will refer to this segmentation system as the *Duke University Speech Time-Aligner*.

| Boundary Misalignment (msec) | | | | |
|---|---|---|---|---|
| $\leq 5$ | $\leq 10$ | $\leq 15$ | $\leq 30$ | $\leq 60$ |
| 47.9 % | 69.9 % | 80.3 % | 92.6 % | 98.4 % |

Table 1: Segmentation accuracy of the Duke University Speech Time-Aligner for (8 kHz) TIMIT database.

## 3. ADDITIVE NOISE COMPENSATION

Many proposed speech enhancement algorithms have been shown to improve speech quality or speech recognition accuracy in additive noise environments. Enhancement algorithms have received little investigation, however, for tasks such as automatic speech segmentation. In general, most methods rely on smoothed estimates of speech and noise characteristics. Since smoothing may reduce the discriminability of phoneme boundaries, algorithms which improve noisy speech recognition may not equally improve segmentation performance in similar noisy environments. In this section we review 4 enhancement methods and subsequently compare each algorithm for noisy speech segmentation.

### 3.1 Generalized Spectral Subtraction (GSS)

In Spectral Subtraction enhancement, an estimate of the clean speech power spectrum, $|\hat{X}(\omega)|^2$, is obtained by subtracting an estimate of the noise power spectrum, $|\hat{N}(\omega)|^2$, from the observed noisy speech power spectrum, $|Y(\omega)|^2$. Generalized Spectral Subtraction (GSS) [4] extends this idea by performing the subtraction in a parametric domain obtained by raising the spectrum to an arbitrary power, $\gamma$. For the $i^{th}$ degraded speech frame, let

$$D_i(\omega) = |Y_i(\omega)|^{2\gamma} - \alpha(SNR_i) \cdot |\hat{N}(\omega)|^{2\gamma}$$

where $\alpha(SNR_i)$ represents a SNR-dependent noise overestimation factor. Using this definition, the GSS estimator is given by,

$$|\hat{X}_i(\omega)|^2 = \begin{cases} D_i(\omega)|^{\frac{1}{\gamma}} & : \; D_i(\omega)^{\frac{1}{\gamma}} > \beta|\hat{N}(\omega)|^2 \\ \beta|\hat{N}(\omega)|^2 & : \; \text{otherwise} \end{cases}$$

where $\beta$ represents a noise flooring factor ($0 < \beta \ll 1$). The incorporation of the noise flooring factor into the rectification operation reduces musical tone artifacts which are present in linear spectral subtraction enhanced speech.

### 3.2 Nonlinear Spectral Subtraction (NSS)

The Nonlinear Spectral Subtraction (NSS) [5] algorithm takes into account frequency-dependent SNR. Using a nonlinear subtractor, the subtraction factor is reduced for spectral components of high SNR and increased for spectral components of low SNR. In addition, the noise model is extended by utilizing both an averaged noise spectrum and an overestimated noise spectrum. The NSS enhancement can be expressed in terms of a filtering operation,

$$|\hat{X}_i(\omega)| = H_i(\omega) \cdot |Y_i(\omega)|,$$

where $H_i(\omega)$ depends on a smoothed estimate of the noisy speech magnitude spectrum, $|\ddot{Y}_i(\omega)|$, and nonlinear subtraction term, $\Phi_i(\omega)$,

$$H_i(\omega) = \frac{|\ddot{Y}_i(\omega)| - \Phi_i(\omega)}{|\ddot{Y}_i(\omega)|}.$$

The subtraction term, $\Phi_i(\omega)$, is given by,

$$\Phi_i(\omega) = \frac{\max\limits_{i-M \leq \tau \leq i}(|N_\tau(\omega)|)}{1 + \gamma\rho_i(\omega)} \quad \text{with} \quad \rho_i(\omega) = \frac{|\ddot{Y}_i(\omega)|}{|\ddot{N}_i(\omega)|}$$

where $\gamma$ is a constant scaling factor dependent on the range of $p_i(\omega)$. Note that for practical purposes, the dynamic range of $\Phi_i(\omega)$ is limited to 1 to 3 times the smoothed noise magnitude estimate (i.e., $|\ddot{N}_i(\omega)| \leq \Phi_i(\omega) \leq 3|\ddot{N}_i(\omega)|$). Like GSS, spectral flooring is performed during enhancement.

### 3.3 Ephraim-Malah MMSE Algorithm

Ephraim and Malah [6] proposed a minimum mean square error (MMSE) estimator for the short-time spectral amplitude component of speech in noise. Here, the speech and noise spectral components are modeled as Gaussian random variables. The algorithm estimates the $k^{th}$ spectral magnitude component using a filter of the form,

$$\begin{aligned} H_i(\omega_k) &= \left(\frac{\sqrt{\pi}}{2}\right)\left(\frac{\sqrt{v_k}}{\gamma_k}\right)\exp\left(-\frac{v_k}{2}\right) \\ &\quad \cdot \left[(1+v_k)\, I_o\left(\frac{v_k}{2}\right) + v_k I_1\left(\frac{v_k}{2}\right)\right] \end{aligned}$$

where $I_o(\cdot)$ and $I_1(\cdot)$ represent modified Bessel functions of the zero and first order. Furthermore, $v_k$ is given by,

$$v_k = \frac{\xi_k}{1+\xi_k}\gamma_k$$

where $\xi_k$ and $\gamma_k$ represent the *a priori* and *a posteriori* signal-to-noise ratios for the $k^{th}$ spectral component. During enhancement processing, a "decision-directed" estimator is used to obtain estimates of $\xi_k$ using knowledge of previously enhanced spectral components.

### 3.4 Constrained Iterative Wiener Filter (AutoLSP)

The Constrained Iterative Wiener Filter (AutoLSP) approach [7] is based upon a two-step maximum *a posteriori* (MAP) estimation of the all-pole speech parameters and noise-free speech. In the first step, a MAP estimation of the clean speech is obtained from the noisy input speech (via Wiener filtering). In the second step, MAP estimation is used to produce the all-pole model parameters given the previous speech estimate. In between MAP estimation steps, spectral constraints are applied in order to (i) ensure stability of the all-pole model, (ii) to ensure that it possesses speech-like characteristics, and (iii) to provide frame-to-frame continuity in vocal tract characteristics. *Inter-frame* constraints are applied to the Line Spectrum Pair (LSP) parameters while *intra-frame* constraints are applied across iterations to the autocorrelation lag sequence.

## 4. ADDITIVE NOISE EXPERIMENTS

### 4.1 Experimental Procedure

Five additive noise environments ranging from highly colored to wideband distortion were considered. These noises include: aircraft cockpit noise (recorded from the cockpit of a C130 transport plane), helicopter fly-by noise, automobile highway noise (recorded in a moving car traveling 95 km/hour), Sun 4/330 workstation cooling fan noise, and white Gaussian noise. Utterances in the core test set of the TIMIT database (1344 sentences $\approx$ 46,000 phoneme boundaries) were corrupted by each noise condition at a level of 10 dB SNR. Waveforms were then aligned with phoneme sequences provided by the TIMIT database. In addition to 4 front-end speech enhancement methods, we also considered the case of HMM models trained with and without cepstral mean subtracted (CMS) parameters. Here, CMS was implemented by subtracting the long-term mean (estimated over the entire utterance) from each mel-cepstral parameter.

It is noteworthy to point out that no rigorous attempt was made to optimize algorithm parameter settings in each noisy environment. Instead, reasonable parameter settings were determined through informal experimentation. Certainly algorithms such as NSS and AutoLSP can be optimized for known environments by appropriate choice of nonlinear subtractor, parameter settings, or terminating iteration count.

### 4.2 Experimental Results

Evaluation results are shown in Table 2. Here, percent of automatically labeled boundaries within a fixed misalignment distance from the actual hand-labeled location is shown. Without compensation (a) a dramatic reduction in segmentation performance is observed (compare with results in Table 1). When models are trained with cepstral mean subtracted parameters (b), segmentation performance improves for all noisy conditions with greatest improvement for white Gaussian noise.

Front-end enhancement processing methods (see descriptions in Table 2 for cases c,d,e,f) were also found to reduce segmentation errors. In general, GSS and NSS provide very similar segmentation improvements. This is not surprising since the NSS algorithm subtracts a variable amount of noise from a smoothed estimate of the speech spectrum during processing. Consequently, the improvements in noise modeling offered by NSS may be outweighed by the smoothing performed on the speech spectrum. In general, however, we see that NSS provides slightly better segmentation performance over the Ephraim-Malah approach in that fewer boundaries are placed outside of the 60 msec interval. Finally, the AutoLSP constrained iterative Wiener filtering approach was found to provide the best overall segmentation performance, although the improvements may be considered only minor in light of the increased computational complexity of the enhancement approach (i.e., iterative Wiener filtering requires roughly an order of magnitude more processing than subtraction based schemes). Still, there is room for considerable improvement considering the segmentation performance in the matched training/testing environment.

COMPARISON OF SPEECH ENHANCEMENT ALGORITHMS FOR IMPROVED AUTOMATIC SEGMENTATION IN NOISE

| | Boundary Misalignment (msec) | | | | |
|---|---|---|---|---|---|
| | $\leq 5$ | $\leq 10$ | $\leq 15$ | $\leq 30$ | $\leq 60$ |
| **Aircraft Cockpit Noise (10 dB SNR)** | | | | | |
| a | 40.4 % | 59.0 % | 68.7 % | 82.5 % | 92.3 % |
| b | 42.8 % | 62.7 % | 72.6 % | 86.0 % | 94.5 % |
| c | 44.0 % | 64.7 % | 74.7 % | 87.8 % | 95.8 % |
| d | 43.2 % | 64.6 % | 75.4 % | 89.1 % | 96.7 % |
| e | 43.3 % | 64.6 % | 75.5 % | 89.0 % | 96.5 % |
| f | 43.6 % | 65.0 % | 75.5 % | 89.1 % | 96.5 % |
| g | 46.1 % | 66.9 % | 77.3 % | 90.6 % | 97.4 % |
| **Helicopter Fly-by Noise (10 dB SNR)** | | | | | |
| a | 34.8 % | 51.7 % | 61.3 % | 75.8 % | 87.4 % |
| b | 37.7 % | 56.6 % | 67.3 % | 82.3 % | 92.2 % |
| c | 39.6 % | 59.3 % | 70.0 % | 84.9 % | 93.9 % |
| d | 39.4 % | 59.0 % | 70.0 % | 85.2 % | 94.3 % |
| e | 39.9 % | 59.6 % | 70.5 % | 85.4 % | 94.3 % |
| f | 39.9 % | 59.8 % | 70.9 % | 86.1 % | 95.0 % |
| g | 43.7 % | 64.4 % | 75.2 % | 89.2 % | 96.9 % |
| **Automobile Highway Noise (10 dB SNR)** | | | | | |
| a | 39.8 % | 58.4 % | 68.5% | 82.7 % | 92.5 % |
| b | 41.1 % | 61.9 % | 72.7% | 86.9 % | 95.2 % |
| c | 42.3 % | 63.6 % | 74.5% | 88.4 % | 96.0 % |
| d | 41.7 % | 62.7 % | 74.2% | 88.4 % | 96.1 % |
| e | 41.5 % | 62.7 % | 74.1% | 88.4 % | 96.0 % |
| f | 43.2 % | 64.8 % | 76.1% | 89.9 % | 97.0 % |
| g | 46.3 % | 67.3 % | 78.0% | 91.4 % | 97.9 % |
| **Sun 4/330 Cooling Fan Noise (10 dB SNR)** | | | | | |
| a | 32.7 % | 48.1 % | 57.0 % | 70.3 % | 82.4 % |
| b | 33.6 % | 50.3 % | 59.8 % | 73.8 % | 85.6 % |
| c | 36.2 % | 54.3 % | 64.3 % | 78.3 % | 89.1 % |
| d | 39.3 % | 58.6 % | 69.2 % | 83.5 % | 93.2 % |
| e | 38.5 % | 57.8 % | 68.4 % | 83.1 % | 92.8 % |
| f | 39.4 % | 58.7 % | 69.6 % | 83.9 % | 93.3 % |
| g | 44.1 % | 65.1 % | 76.0 % | 89.8 % | 97.0 % |
| **White Gaussian Noise (10 dB SNR)** | | | | | |
| a | 29.4 % | 43.5 % | 51.6 % | 64.9 % | 77.0 % |
| b | 35.6 % | 52.2 % | 61.2 % | 75.1 % | 86.4 % |
| c | 38.5 % | 55.7 % | 64.9 % | 78.6 % | 89.0 % |
| d | 39.3 % | 57.8 % | 67.8 % | 82.2 % | 92.3 % |
| e | 38.9 % | 57.1 % | 67.2 % | 81.5 % | 91.6 % |
| f | 39.4 % | 58.1 % | 68.4 % | 83.0 % | 92.6 % |
| g | 43.7 % | 62.8 % | 72.9 % | 86.7 % | 95.6 % |

Table 2: Speech segmentation accuracy for additive noise environments for (a) no compensation, (b) Cepstral Mean Subtraction (CMS), (c) CMS and Ephraim-Malah MMSE enhancement, (d) CMS and Generalized Spectral Subtraction, (e) CMS and Nonlinear Spectral Subtraction, (f) CMS and AutoLSP Constrained Iterative Wiener filter, (g) models retrained in noisy environment.

## 5. TELEPHONE CHANNEL COMPENSATION

### 5.1 Experimental Procedure

An investigation of speech segmentation performance in noisy telephone channel environments was also conducted. The NTIMIT [8] database (TIMIT transmitted over telephone channels) and CTIMIT [9] database (TIMIT transmitted over cellular channels) were used for evaluation. Segmentation performance was considered for several front-end processing scenarios including: no compensation, cepstral mean subtraction, CMS + GSS or AutoLSP enhancement, CMS + AutoLSP with models trained via a wireline simulator. For the latter case, TIMIT training speech was passed through a wireline channel simulator [10] in order to simulate telephone channel degradations. In particular, we used the Continental United States mid-quality voice (CMV) channel. Additionally, we compared results obtained by retraining the models in the matched environment. For each experiment, HMM models were trained using mel-scale filter banks spanning the telephone bandwidth of 300 to 3300 Hz. Segmentation performance was determined as described in Section 4.1.

### 5.2 Experimental Results

Results for telephone based speech segmentation are shown in Table 3. For both NTIMIT and CTIMIT databases, cepstral mean subtraction was found to provide the greatest single improvement in segmentation performance. Second, we note that AutoLSP provides additional performance over CMS for both CTIMIT and NTIMIT speech. This improvement is more significant for CTIMIT since audible background noise distortions can be heard more frequently on the CTIMIT database compared with the NTIMIT database. Finally, the use of models trained by passing clean TIMIT speech through a wireline channel simulator was also found to be beneficial. Note, however, that this does not take into account distortions introduced by handset variabilities.

## 6. CONCLUSIONS

In this paper we have compared several previously proposed speech enhancement algorithms for application to automatic segmentation of speech recorded in non-ideal environments. Although, these algorithms have been shown to provide improved speech quality and speech recognition performance for noisy environments, many methods rely on smoothed estimates of the noise or speech characteristics during enhancement processing. Consequently, algorithms which improve speech recognition performance may not equally improve noisy speech segmentation. Overall, we found that the AutoLSP Constrained Iterative Enhancement algorithm provided the greatest improvement in additive noise. However, there is still considerable mismatch between segmentation accuracy in matched environments. For telephone based speech segmentation, cepstral mean subtraction was found to significantly improve segmentation performance. Additional incorporation of front-end enhancement and retraining models using a wireline simulator was also found to provide improved robustness to telephone channel degradations.

| | Boundary Misalignment (msec) | | | | |
|---|---|---|---|---|---|
| | $\leq 5$ | $\leq 10$ | $\leq 15$ | $\leq 30$ | $\leq 60$ |
| **NTIMIT Telephone Database** | | | | | |
| a | 25.1 % | 40.3 % | 50.0 % | 62.9 % | 72.8 % |
| b | 32.1 % | 52.4 % | 65.5 % | 81.2 % | 90.9 % |
| c | 32.5 % | 52.7 % | 65.7 % | 81.7 % | 91.3 % |
| d | 33.5 % | 54.4 % | 67.5 % | 83.0 % | 92.3 % |
| e | 32.3 % | 54.3 % | 68.6 % | 85.3 % | 94.3 % |
| f | 41.4 % | 63.8 % | 75.9 % | 90.3 % | 97.4 % |
| **CTIMIT Cellular Telephone Database** | | | | | |
| a | 17.7 % | 26.0 % | 31.1 % | 39.6 % | 46.6 % |
| b | 29.9 % | 45.6 % | 54.5 % | 68.7 % | 79.8 % |
| c | 31.2 % | 46.9 % | 56.5 % | 71.7 % | 82.6 % |
| d | 31.5 % | 47.9 % | 57.4 % | 72.5 % | 83.4 % |
| e | 31.4 % | 47.6 % | 57.4 % | 72.6 % | 83.6 % |
| f | 34.5 % | 52.5 % | 63.7 % | 80.0 % | 90.8 % |

Table 3: Comparison of front-end processing schemes for segmentation of telephone and cellular telephone speech. Methods shown include (a) no compensation, (b) CMS, (c) CMS + GSS, (d) CMS + AutoLSP, (e) Wireline simulator models + CMS + AutoLSP, (f) matched environment.

## REFERENCES

[1] A. Lijolie, M. Riley, "Automatic Segmentation and Labeling of Speech," *Proc. IEEE ICASSP,* pp. 473-476, 1991.

[2] F. Brugnara, D. Falavigna, M. Omologo, "Automatic Segmentation and labeling of speech based on Hidden Markov Models," *Speech Communication,* vol. 12, pp. 357-370, 1993.

[3] E. Giachin, A. Rosenberg, C.H. Lee, "Word Juncture Modeling using phonological rules for HMM-based continuous speech recognition," *Computer Speech and Language,* vol. 5, pp. 155-168, 1991.

[4] M. Berouti, R. Shwartz, J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *Proc. IEEE ICASSP,* pp. 208-211, 1979.

[5] P. Lockwood, J. Boudy, "Experiments with a Nonlinear Spectral Subtractor (NSS), Hidden Markov Models and the projection, for Robust Speech Recognition in Cars," *Speech Communication,* vol. 11, pp. 215-228, 1992.

[6] Y. Ephraim, D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. ASSP,* vol. ASSP-32, no. 6, pp. 1109-1121, 1984.

[7] J. Hansen, M. Clements, "Constrained Iterative Speech Enhancement with Application to Speech Recognition," *IEEE Trans. on Signal Processing,* vol. 39, no. 4, pp. 795-805, 1991.

[8] C. Jankowski, A. Kalyanswamy, S. Basson, J. Spitz, "NTIMIT: A phonetically balanced, Continuous Speech, Telephone Bandwidth Speech Database," *Proc. IEEE ICASSP,* pp. 109-112, 1990.

[9] K. Brown, E. George, "CTIMIT: A Speech Corpus for the Cellular Environment with Applications to Automatic Speech Recognition," *Proc. IEEE ICASSP,* pp. 105-108, 1995.

[10] J. Kupin, "A Wireline Simulator [Software]," CCR-P, April 1993.

[11] L. Neumeyer, V. Digalakis, M. Wientraub, "Training Issues and Channel Equalization Techniques for the Construction of Telephone Acoustic Models Using a High-Quality Speech Corpus," *IEEE Trans. on Speech and Audio Processing,* vol. 2, no. 4, pp. 590-597, Oct. 1994.

[12] B. Petek, O. Anderson, P. Dalsgaard, "On the Robust Automatic Segmentation of Spontaneous Speech," *Proc. ICSLP,* pp. 913-916, 1996.