Research Article

Speech Waveform Compression Using Robust Adaptive Voice Activity Detection for Nonstationary Noise

Waheeduddin Q. Syed and Hsiao-Chun Wu

Communications and Signal Processing Laboratory, Department of Electrical and Computer Engineering, Louisiana State University, Baton Rouge, LA 70803, USA

Correspondence should be addressed to Hsiao-Chun Wu, wu@ece.lsu.edu

Received 8 June 2007; Revised 17 December 2007; Accepted 18 February 2008

Recommended by Dan Chazan

The voice activity detection (VAD) is crucial in all kinds of speech applications. However, almost all existing VAD algorithms suffer from the nonstationarity of both speech and noise. To combat this difficulty, we propose a new voice activity detector, which is based on the Mel-energy features and an adaptive threshold related to the signal-to-noise ratio (SNR) estimates. In this paper, we first justify the robustness of the Bayes classifier using the Mel-energy features over that using the Fourier spectral features in various noise environments. Then, we design an algorithm using the dynamic Mel-energy estimator and the adaptive threshold, which depends on the SNR estimates. In addition, a realignment scheme is incorporated to correct the sparse-and-spurious noise estimates. Numerous simulations are carried out to evaluate the performance of our proposed VAD method and the comparisons are made with a couple of existing representative schemes, namely, the VAD using the likelihood ratio test with Fourier spectral energy features and that based on the enhanced time-frequency parameters. Three types of noises, namely, white noise (stationary), babble noise (nonstationary), and vehicular noise (nonstationary) were artificially added by the computer for our experiments. As a result, our proposed VAD algorithm significantly outperforms other existing methods as illustrated by the corresponding receiver operating characteristics (ROC) curves. Finally, we demonstrate one of the major applications, namely, speech waveform compression associated with our new robust VAD scheme and quantify the effectiveness in terms of compression efficiency.

Copyright © 2008 W. Q. Syed and H.-C. Wu. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. INTRODUCTION

Nowadays, speech processing techniques can be applied in a wide variety of communication devices such as cellular handsets, internet search machines, and call-in telephony services. Despite its constant growth, voice activity detection, one of important speech processing techniques, is still intriguing to many researchers [1]. A voice activity detector is a preprocessing component for speech recognition systems, isolated word boundary detection systems, cellular phones, and speech enhancement systems. Voice activity detection (VAD) algorithm is designed to distinguish the speech from the background noise among short-time frames. The importance of the VAD system to the speech processing applications can be easily found in the existing literature. In the early developed VAD algorithms, the features were extracted from the short-time energy, zero-crossing rates [2], linear predictive coding coefficients [3], and cepstral coefficients [4]. Recently, the Mel-energy features [5], the wavelet transforms [6], the correlation coefficients [7], and the likelihood ratios [8, 9] have been adopted as the primary features for the VAD techniques. Among all of the adopted features for VAD, the Mel-spectra have been shown to be very promising in the previous VAD method using the threshold based on the minimum subband Mel energy [5]. According to several experiments in [5], the Mel-spectral features would lead to the most robust VAD performance compared with almost all other features. However, the comprehensive studies associated with the Mel-spectral features for VAD cannot be found in the existing literature.

In most of the aforementioned techniques, the feature extraction is followed by the threshold detection. In the realistic environments, there exist nonstationary noises such as babble noise and vehicular noise. Therefore, the static thresholds, which can only depend on the information extracted from the first few frames, would cause numerous
classification errors [7]. Hence, dynamic or adaptive thresholds were proposed to combat the problem of nonstationary noises [5]. Nevertheless, how to appropriately adjust the threshold dynamically is still very challenging up to now [5, 7]. In this paper, we propose a new adaptive threshold, which depends on the signal-to-noise ratio (SNR) estimates and results from the dynamical speech and noise information. Consequently, it can lead to the much better VAD performance in the presence of nonstationary noise. Furthermore, we extend our new VAD technique for the application of the speech waveform compression, which can be used in the voice communications and storage [10].

In this paper, we first justify the advantages of the Mel-energy features via the Bayes hypothesis analysis instead of the psychophysical conjectures in the existing literature [5]. We then propose a new robust VAD algorithm, which is based on the Mel-energy features and the adaptive threshold detection. Such a dynamical threshold can be derived from the SNR estimates in [11]. The rest of this paper is organized as follows. The time-frequency features of the speech signals are studied and analyzed in Section 2. In Section 3, we introduce our new robust VAD scheme using the Mel-energy features and the SNR-based adaptive threshold. The simulation results to demonstrate the effectiveness of our new VAD algorithm and the speech compression performance are both presented in Section 4. The concluding remarks will be finally drawn in Section 5.

2. COMPARATIVE STUDIES FOR TIME-FREQUENCY FEATURES

The features play a key role in all voice activity detectors. The ambiguity due to the unreliable features, especially in the conditions of low SNRs and/or nonstationary background noises, will cause the misdetection very often [1]. According to a few simulations in [5], robust speech detection can be achieved using the Mel-spectral features compared to the Fourier spectral features. The existing literature provides the explanation simply based on the auditory psychophysics that the human ears perceive acoustic waves along the nonlinear scale in the frequency domain, which forms the Mel filter bank [5]. In addition, the dimensionality reduction, but not in the tradeoff of performance, can be achieved for speech detection and recognition using the Mel-spectral features. In the subsequent sections, we formulate the Fourier spectral and the Mel-spectral features for VAD. We compare the VAD performances using these two features via the Bayes hypothesis analysis accordingly to show the effectiveness of the Mel-spectral features.

2.1. Fourier spectral features

Fourier spectral features are obtained from the short-time Fourier transform. A primary Fourier spectral feature, short-time Fourier energy $|x_{\text{freq}}[n, k]|^2$, is defined as [12]

$$
|x_{\text{freq}}[n, k]|^2 \equiv \left| \sum_{m=0}^{N-1} x(n-m)w(m)e^{-j\frac{2\pi km}{N}} \right|^2,
$$

where $x(n)$ is the discrete-time speech signal, $w(m)$ is the window size, and $N$ is the window size. According to (1), it is noted that $|x_{\text{freq}}[n, k]|^2$ is a double-indexed function with time index $n$ and frequency index $k$. Usually, the short-time framed Fourier energy $E_{\text{FT}}[n', k]$ will be collected for $n = 0, \Delta n, 2\Delta n, \ldots, n'\Delta n, \ldots$ and $n' \in Z^+ \cup \{0\}$, such that

$$
E_{\text{FT}}[n', k] \equiv \left| x_{\text{freq}}[n'\Delta n, k] \right|^2, \text{ for } n' \in Z^+ \cup \{0\},
$$

where $\Delta n > 0$ is the frame advance step size.

2.2. Mel-spectral features

The Mel-spectral features can be acquired through the weighted Fourier spectral features via the Mel filter bank, which is a uniformly spaced filter bank on a nonlinearly wrapped frequency scale, known as the Mel scale, as illustrated in Figure 1. The relationship between the Mel-scale frequency $f_{\text{mel}}$ and the conventional frequency $f_{\text{son}}$ (in Hz) is given by [5]

$$
f_{\text{mel}} = 2595 \log\left(1 + \frac{f_{\text{son}}}{700}\right).
$$

Without loss of generality, we choose a set of 20-band Mel filters throughout this paper as illustrated in Figure 1. As depicted in Figure 1, the squared magnitude response of the $i$th Mel filter,

$$
|H_{\text{mel}}(i, k)|^2, \quad 0 \leq k \leq N - 1, \quad 1 \leq i \leq 20,
$$

Figure 1: A Mel filter bank composed of the triangular bandpass filters each with a bandwidth and the spacing in accordance with the Mel scale in the frequency domain.
are the two corresponding hypotheses. Let the two a priori

\[ \omega_n \] and \( \omega_s \) specify the individual weighting factor for the \( k \)th frequency component of the Fourier spectra [5]. According to (2) and (3), the short-time framed Mel energy \( E_{mel}(n', i) \) is given by

\[
E_{mel}(n', i) = \sum_{k=0}^{N-1} E_{TT}[n', k] \left| H_{mel}(i, k) \right|^2, \quad 1 \leq i \leq 20,
\]

where \( i \) is the Mel-filter index. It is noted that the frequency dimensionality of \( E_{mel}(n', i) \) is reduced from \( N \) of \( E_{TT}[n', k] \) to 20.

**2.3. Comparative studies between Fourier and Mel-spectral analyses for speech**

In this section, we will provide the simulation results for the speech/noise classification to justify the advantage of the Mel-energy features over the Fourier energy features extracted on frame-by-frame basis. We establish an optimal Bayes classifier in [13] to evaluate the effectiveness of the extracted features under the assumption that the entire set of feature vectors are drawn from the multidimensional Gaussian processes. The general framework of the Bayes classifiers is shown in Figure 2. Provided speech data, the framed Fourier and Mel energies are acquired to establish the corresponding Bayes classifiers and the ground truth (the true speech/noise frame labels), which are applied to determine the optimal threshold. Then, the outcomes of each classifier will be compared with the ground truth.

The Bayes classifiers can be constructed as follows. Each feature vector is of dimension \( d \) (\( d = N \) for Fourier spectral features, \( d = 20 \) for Mel-spectral features). For a feature vector in an arbitrary frame, let us denote it as \( \tilde{x} \in \mathbb{R}^d \). Thus, the speech/noise frame classification becomes a binary Bayes hypothesis test, where \( H_s : \tilde{x} \in \omega_s \) (\( \tilde{x} \) is extracted from a frame in the presence of both speech and noise) and \( H_n : \tilde{x} \in \omega_n \) (\( \tilde{x} \) is extracted from a frame in the presence of noise only) are the two corresponding hypotheses. Let the two a priori

probabilities be \( P(\omega_s) \) and \( P(\omega_n) \), respectively. Then, the a posteriori probabilities are given by [13]

\[
P(\omega_s \mid \tilde{x}) = \frac{P(\tilde{x} \mid \omega_s)P(\omega_s)}{P(\tilde{x})},
\]

\[
P(\omega_n \mid \tilde{x}) = \frac{P(\tilde{x} \mid \omega_n)P(\omega_n)}{P(\tilde{x})},
\]

where \( P(\tilde{x}) = P(\tilde{x} \mid \omega_s)P(\omega_s) + P(\tilde{x} \mid \omega_n)P(\omega_n) \) is a common factor associated with the nonparametric probability. The conditional probabilities are given by

\[
P(\tilde{x} \mid \omega_s) = \frac{1}{\sqrt{(2\pi)^d|\tilde{\Sigma}_s|}} \exp \left[ -\frac{1}{2} (\tilde{x} - \tilde{\mu}_s)^T \tilde{\Sigma}_s^{-1} (\tilde{x} - \tilde{\mu}_s) \right],
\]

\[
P(\tilde{x} \mid \omega_n) = \frac{1}{\sqrt{(2\pi)^d|\tilde{\Sigma}_n|}} \exp \left[ -\frac{1}{2} (\tilde{x} - \tilde{\mu}_n)^T \tilde{\Sigma}_n^{-1} (\tilde{x} - \tilde{\mu}_n) \right],
\]

where \( \tilde{\mu}_s, \tilde{\mu}_n \) are the mean vectors and \( \tilde{\Sigma}_s, \tilde{\Sigma}_n \) are the covariance matrices associated with the extracted feature vectors \( \tilde{x} \) for Hypotheses \( H_s \) and \( H_n \), respectively. According to (6), the Bayes classifier depends on \( P(\tilde{x} \mid \omega_s), P(\tilde{x} \mid \omega_n), P(\omega_s) \), and \( P(\omega_n) \) only. The ground truth can be utilized to determine \( P(\omega_s) \) and \( P(\omega_n) \). In addition, the ground truth can also be utilized to determine the logarithms of the conditional probabilities as well, such that

\[
\log[P(\tilde{x} \mid \omega_s)] = -\frac{d}{2} \log(2\pi) - \frac{1}{2} \log(|\tilde{\Sigma}_s|)
\]

\[
-\frac{1}{2} (\tilde{x} - \tilde{\mu}_s)^T \tilde{\Sigma}_s^{-1} (\tilde{x} - \tilde{\mu}_s),
\]

\[
\log[P(\tilde{x} \mid \omega_n)] = -\frac{d}{2} \log(2\pi) - \frac{1}{2} \log(|\tilde{\Sigma}_n|)
\]

\[
-\frac{1}{2} (\tilde{x} - \tilde{\mu}_n)^T \tilde{\Sigma}_n^{-1} (\tilde{x} - \tilde{\mu}_n).
\]

According to (6)–(9), the Bayes decision rule is given by

\[
\tilde{x} \in \omega_s \quad \Gamma(\omega_s \mid \tilde{x}) > \Gamma(\omega_n \mid \tilde{x}),
\]

\[
\tilde{x} \in \omega_n \quad \Gamma(\omega_n \mid \tilde{x}) < \Gamma(\omega_s \mid \tilde{x}),
\]

where the discriminant functions are defined as

\[
\Gamma(\omega_s \mid \tilde{x}) = \log[P(\omega_s)] - \frac{1}{2} \log(|\tilde{\Sigma}_s|)
\]

\[
-\frac{1}{2} (\tilde{x} - \tilde{\mu}_s)^T \tilde{\Sigma}_s^{-1} (\tilde{x} - \tilde{\mu}_s),
\]

\[
\Gamma(\omega_n \mid \tilde{x}) = \log[P(\omega_n)] - \frac{1}{2} \log(|\tilde{\Sigma}_n|)
\]

\[
-\frac{1}{2} (\tilde{x} - \tilde{\mu}_n)^T \tilde{\Sigma}_n^{-1} (\tilde{x} - \tilde{\mu}_n).
\]

For quantification convenience, we artificially add the clean speech with noise (SNR = 5 dB) and carry out the Bayes
classifiers as given by (10). The speech data are randomly picked from the TIMIT database (three male and three female speakers) [14], while the white and the babble noises are taken from the NOISEX-92 database [15]. The ground truth comes from the speech frame labels specified in the TIMIT database. The outcomes of the Bayes classifiers, in terms of the percentages of classification errors for both Mel-spectral features and Fourier spectral features, are presented in Table 1. According to Table 1, the Mel-spectral features lead to the much better speech/noise frame detection performance than the Fourier spectral features in both white and babble noises.

### 3. NOVEL VOICE DETECTION USING MEL-SPECTRA AND ADAPTIVE THERSHOLDS

In this section, we introduce the feature extraction and the adaptive threshold determination and thereby, propose the new Mel energy-based adaptive-threshold voice activity detector (ME VAD). Feature extraction and threshold adaptation are presented in Sections 3.1 and 3.2, respectively. A new realignment mechanism to improve the classification accuracy is also addressed in Section 3.3.

#### 3.1. Mel-energy feature extraction

Consider a time-domain noisy speech signal, which is divided into the overlapping frames each of 256 samples. In addition, each frame is weighted by a Hamming window of the same size. Initially, the short-time framed Fourier energy \( E_{\text{FT}}\{n', k\} \) can be determined using (2). The short-time framed Mel-energy \( E_{\text{mel}}\{n', i\} \) for the \( i \)th frequency band in the \( n' \)th frame can be calculated using (5). Thus, for the \( n' \)th frame, the robust energy indicator \( I(\ n') \) can be calculated as [5]

\[
I(n') = \sum_{i=1}^{20} E_{\text{mel}}(n', i).
\]

For example, \( I(n') \) is depicted for a speech signal corrupted by the babble noise for SNR = 5 dB in Figure 3.

#### 3.2. Adaptive threshold determination

We propose a new “two-threshold” scheme for a better speech/noise classification. Since the first \( N_{\text{noise}} \) frames consist of noise only, we can determine the first threshold or the a priori threshold \( \eta_{\text{apr}}(n') \), for the frame \( n' \), as

\[
\eta_{\text{apr}}(n') = \min \left[ 1.2 E_n, \frac{E_{\text{max}}(n') + E_n}{2} \right],
\]

where \( E_n = (1/N_{\text{noise}}) \sum_{m=n-1}^{n-1} I(m) \) denotes the average energy of the first few noise frames, \( E_{\text{max}}(n') = \max_{1 \leq m \leq n' \leq n} \{ I(m) \} \). Accordingly, the a priori decision rule is given by

\[
I(n') \begin{cases} \text{speech} & > \eta_{\text{apr}}(n') \\ \text{noise} & < \eta_{\text{apr}}(n') \end{cases}
\]

After the first \( N_{\text{noise}} \) noise frames, the classified noise energy indicator \( N(n') = I(n') \), if the \( n' \)th frame is the noise frame according to (14), is stored in a noise energy buffer, while \( S(n') = I(n') \) is stored in a speech energy buffer if such a frame is the speech frame instead. Once the sufficient collection of both \( N(n') \) and \( S(n') \) is available, the SNR estimation can be achieved using [11].

Next, we introduce the procedure for estimating temporal SNR and the a posteriori threshold \( \eta_{\text{aps}}(n') \) as follows. The temporal speech energy estimate \( \hat{S}(n') \) can be obtained as

\[
\hat{S}(n') = \sum_{m=n' - B_s + 1}^{n'} S(m) - \sum_{m=n' - B_n + 1}^{n'} N(m),
\]

where \( B_s \) and \( B_n \) specify the speech buffer and noise buffer sizes, respectively. Thereby, the temporal SNR can be calculated as

\[
\text{SNR}(n') = 10 \log_{10} \left[ \frac{\hat{S}(n')}{\sum_{m=n' - B_n + 1}^{n'} N(m)} \right].
\]

According to (16), we can estimate the temporal noise energy estimate \( \hat{N}(n') \) as

\[
\hat{N}(n') = \frac{E_{\text{max}}(N_{\text{noise}})}{1 + \gamma \text{SNR}(n')},
\]

| Types of noise          | Male speakers |            | Female speakers |            |
|------------------------|---------------|------------|-----------------|------------|
|                        | Data samples  | Mel-spectral features | Data samples  | Mel-spectral features |
|                        |               |   |               |               |
| White noise (SNR = 5 dB)| Speaker1      | 14.13%   | Speaker1        | 17.40%      |
|                        | Speaker2      | 23.85%   | Speaker2        | 24.92%      |
|                        | Speaker3      | 12.18%   | Speaker3        | 31.35%      |
|                        | Average       | 16.72%   | Average         | 24.55%      |
| Babble noise (SNR = 5 dB)| Speaker1    | 13.66%   | Speaker1        | 17.26%      |
|                        | Speaker2      | 19.85%   | Speaker2        | 23.68%      |
|                        | Speaker3      | 15.90%   | Speaker3        | 32.62%      |
|                        | Average       | 16.47%   | Average         | 24.52%      |

### Table 1: Percentages of classification errors using the bayes classifiers for white noise and babble noise.
where \( y \) is the control parameter. Thus, the a posteriori threshold \( \eta_{aps}(n') \) can be determined as
\[
\eta_{aps}(n') \equiv \min\left[ 1.2\hat{N}(n'), \frac{E_{\text{max}}(n') + \hat{N}(n')}{2} \right]. \tag{18}
\]

The a posteriori speech/noise frame classification can be achieved using
\[
I(n') > \eta_{aps}(n') \quad \text{speech}
\]
\[
I(n') < \eta_{aps}(n') \quad \text{noise} \tag{19}
\]

According to (13) and (18), the two thresholds are dynamically adapted and therefore, they can track the nonstationary. Finally, these speech/noise frame labels are sent through a realignment mechanism to remove the sparse occurrence of speech/noise as described in the next section.

### 3.3. Realignment

Maleh and Kabal concluded that the hangover schemes were not effective in correcting isolated VAD errors (i.e., a speech frame among a sequence of noise frames or vice versa) and hence they proposed an isolated error correction mechanism (IECM) [16]. Our realignment mechanism is simply based on the similar approach using the majority voting to forcefully reassign the labels resulting from (19) consistent with the majority every 5 to 7 successive frames. The complete ME VAD algorithm is presented in Figure 4.

### 4. SIMULATIONS

Here, we provide the performance comparison for the different noise environments at various noise levels. Three major VAD algorithms are carried out for simulations, namely, our ME VAD method, statistical model-based voice activity detector (SM VAD) [8] and enhanced time-frequency-based robust word boundary detection algorithm (ETF VAD) [5]. The signals in experiments were drawn from the TIMIT database [14] and corrupted by the nonstationary babble and vehicular noises from the NOISEX-92 database [15]. The sampling frequency was set as 16 kHz. Thirty six data files were generated by the computer; each data file under test consisted of three different speech utterances from the same speaker concatenated together with pauses in between and added with the aforementioned noise samples for SNR = 5 dB and SNR = 15 dB. The simulation results will be given next.

#### 4.1. Receiver operating characteristics (ROC) for VAD

The receiver operating characteristics (ROC) curve of a detector is often utilized to illustrate the overall detection performance [17]. The corresponding ROC curves for the aforementioned VAD algorithms are depicted in Figures 5 and 6 for SNR = 5 dB and SNR = 15 dB, respectively. It can be observed that the performance of ME VAD (our algorithm) is almost always better than the ETF VAD and SM VAD methods for the nonstationary babble and vehicular noises. For the critical false detection rates of 2% and 4%, the correct detection rates are tabulated in Tables 2 and 3 for SNR=5 dB and SNR = 15 dB, respectively. According to Figures 5 and 6, Tables 2 and 3, our ME VAD algorithm is superior to the other exiting methods at SNR = 5 dB and comparable to the other existing methods at SNR = 15 dB.
4.2. Speech waveform compression

In addition, we develop a speech waveform compression scheme using the ME VAD outcomes. This algorithm identifies the beginnings and the ends of the pauses in a speech signal and then deletes the detected noise frames for waveform compression. The speech reconstruction (decompression) can be performed by reinserting the pauses at a desired noise level [18]. The performance of any compression scheme is generally measured in terms of compression efficiency and playback quality [19]. However, the compression efficiency measure might be misleading since it dramatically varies among data and strictly depends on the durations of the silence periods. Therefore, in this paper, we define a new compression efficiency measure $C_d$ as

$$C_d = \frac{\text{The number of detected noise frames}}{\text{The number of total frames}} \times \%.$$

(20)
On the other hand, the nature of the original speech data can be characterized as the actual noise percentage measure $C_a$, that is,

$$C_a = \frac{\text{The number of actual noise frames}}{\text{The number of total frames}} \times 100\%.$$  
(21)

The compression efficiency curves for different speech data files (added with babble noise at SNR = 5 dB) as depicted in Figure 7 illustrate the relationship between $C_d$ and $C_a$ from the ground truth. It is obvious that the optimal compression is achieved only when $C_d = C_a$. Thus, the closer the $C_d - C_a$ curve for each scheme to the straight line $C_a = C_a$ as illustrated as “Optimum” in Figure 7, the better the compression performance. According to Figure 7, our ME VAD-based speech compression method significantly outperforms others.

5. CONCLUSION

In this paper, we study the advantage of speech/noise detection using the Mel-spectral features over the Fourier spectral features via the Bayes hypothesis analysis. Then, we design a robust speech/noise voice activity detection algorithm using the adaptive a priori and a posteriori thresholds incorporated with a realignment scheme. Moreover, we can extend this new voice activity detection method to establish a new speech waveform compressor for voice communications. Simulation results show that our new voice activity detection algorithm and the corresponding speech compression scheme greatly outperform other existing methods, especially in the low signal-to-noise ratio environments of babble and vehicular noises.

Figure 7: Compression efficiency $C_d - C_a$ curves for the four schemes using the ME VAD, ETF VAD, SM VAD, and ideal case (optimum).

ACKNOWLEDGMENTS

This research work has been supported by Information Technology Research Award for National Priorities from the National Science Foundation (NSF-ECS 0426644), Research Initiation Grant from the Southeastern Center for Electrical Engineering Education, Research Enhancement Award from the Louisiana-NASA Space Consortium, NSF-Louisiana EPSCOR Pilot Fund, and Faculty Research Grant from the Louisiana State University.
REFERENCES

[1] M. Marzinzik and B. Kollmeier, “Speech pause detection for noise spectrum estimation by tracking power envelope dynamics,” IEEE Transactions on Speech and Audio Processing, vol. 10, no. 2, pp. 109–118, 2002.

[2] M. Hahn and C. K. Park, “An improved speech detection algorithm for isolated Korean utterances,” in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP ’92), vol. 1, pp. 525–528, San Francisco, Calif, USA, March 1992.

[3] L. R. Rabine and M. R. Sambur, “Voiced-unvoiced-silence detection using the Itakura LPC distance measure,” in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP ’77), vol. 2, pp. 323–326, Hartford, Conn, USA, May 1977.

[4] J. A. Haigh and J. S. Mason, “Robust voice activity detection using cepstral features,” in Proceedings of the IEEE Region 10 Conference on Computer, Communication, Control, and Power Engineering (TENCON ’93), vol. 3, pp. 321–324, Beijing, China, October 1993.

[5] C.-T. Lin, J.-Y. Lin, and G.-D. Wu, “A robust word boundary detection algorithm for variable noise-level environment in cars,” IEEE Transactions on Intelligent Transportation Systems, vol. 3, no. 1, pp. 89–100, 2002.

[6] S.-H. Chen and J.-F. Wang, “A wavelet-based voice activity detection algorithm in noisy environments,” in Proceedings of 9th International Conference on Electronics, Circuits, and Systems (ICECS ’02), vol. 3, pp. 995–998, Dubrovnik, Croatia, September 2002.

[7] A. Craciun and M. Gabrea, “Correlation coefficient-based voice activity detector algorithm,” in Proceedings of the IEEE Canadian Conference on Electrical and Computer Engineering (CCECE ’04), vol. 3, pp. 1789–1792, Niagara Falls, Canada, May 2004.

[8] J. Sohn, N. S. Kim, and W. Sung, “A statistical model-based voice activity detection,” IEEE Signal Processing Letters, vol. 6, no. 1, pp. 1–3, 1999.

[9] J. Sohn and W. Sung, “A voice activity detector employing soft decision based noise spectrum adaptation,” in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP ’98), vol. 1, pp. 365–368, Seattle, Wash, USA, May 1998.

[10] A. Benyassine, E. Shlomot, H.-Y. Su, D. Massaloux, C. Lamblin, and J.-P. Petit, “ITU-T Recommendation G.729 Annex B: a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications,” IEEE Communications Magazine, vol. 35, no. 9, pp. 64–73, 1997.

[11] 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, vol. 5, no. 1, pp. 11–21, 1997.

[12] S. K. Mitra, Digital Signal Processing: A Computer Based Approach, Mc-Graw Hill, Boston, Mass, USA, 2nd edition, 2001.

[13] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, John Wiley & Sons, New York, NY, USA, 2nd edition, 2002.

[14] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, “DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM,” NIST, 1986.

[15] A. 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., Defense Evaluation and Research Agency (DERA), Speech Research Unit, Malvern, UK, 1992.