

A ROBUST ENDPOINT DETECTION OF SPEECH FOR NOISY ENVIRONMENTS WITH APPLICATION TO AUTOMATIC SPEECH RECOGNITION

Sahar E. Bou-Ghazale and Khaled Assaleh

Conexant Systems Inc.
Wireless Communications Division
4311 Jamboree Road, K02-851
Newport Beach, CA 92660, USA
sahar.bou-ghazale@conexant.com khaled.assaleh@conexant.com

ABSTRACT

We propose a new approach for classifying speech vs. non-speech, which proves to significantly improve speech recognition performance under noise. The proposed algorithm relies on the energy and spectral characteristics of the signal and applies a 3-level two-dimensional thresholding to determine whether an input frame is speech or non-speech. The algorithm runs in real-time, and offers better immunity to background noise, and to background speech than traditional energy-based word boundary detection. The performance of the endpoint detector is reported here in terms of improvements in speaker-independent (SI) and speaker-dependent (SD) recognition performance using 5 different simulated noise conditions and various signal-to-noise ratios (SNR). The proposed endpoint detection of speech improves the SD recognition accuracy by 24% for office noise, and reduces the false rejection rates for both SI and SD by 45% for babble noise and lobby noise.

1. INTRODUCTION

Accurate endpoint detection is crucial for good speech recognition accuracy. Energy-based endpoint detection algorithms are computationally efficient and perform adequately in quiet conditions but they have little to no immunity to background noise. Offering immunity to background noise becomes critical when using automatic speech recognition in real-life applications such as in a car environment, on the street, in a lobby, etc. In the presence of background noise, an energy-based endpointer often fails to determine the accurate endpoints of a speech utterance since it relies on a single measure for classifying speech vs. silence. By relying on energy alone and if the background is noisy, it is possible to (1) miss the leading or trailing low-energy sounds such as fricatives, (2) classify clicks, pops and background noises as part of speech due to their high energy content, and (3) falsely classify background noise as speech while missing the actual speech. These types of errors result in inaccurate speech boundaries that often lead to high false rejection rates, which in turn reflect negatively on the overall performance of the recognition system. Moreover, endpoint detection is crucial in the training phase of SD systems. An

erroneous determination of endpoints can result in poor word models that have a severe effect on the recognition phase.

Our goal is to devise an endpoint detection that runs in real-time, i.e., requires no additional delay, requires low memory and computational resources, and minimizes the number of errors common to most endpoint detection algorithms in a noisy environment. In particular, it is desired to minimize the false classifications of clicks, pops, lip smacks, and background noises as part of speech and to limit the chopping of low-energy sounds such as fricatives.

In energy-based end point detection algorithms, most clicks and pops are classified as speech due to their high energy content. However, the spectral characteristics of a click differ greatly from those of speech. Therefore, incorporation of spectral information resolves most of the confusion related to misclassifying pops and clicks as speech. In order to incorporate spectral information, and in order to keep the computational resources of the integrated system at a minimum, it is advantageous to design an endpoint detection algorithm around a feature set that is already being used by the recognition system. It is also essential not to introduce any additional delays beyond those required for energy based end point detection. Utilizing cepstral features in the endpoint detection algorithm seems to be a natural choice since they are the basis of numerous speech recognition and verification systems.

Aside from being the basis of most recognition systems, cepstral parameters possess many desirable attributes that make them ideal for our purpose. These include independence from absolute signal levels, robustness towards slow signal amplitude variations, and relative robustness to noise levels [6]. Moreover, the cepstral features coming from stationary noise are known to have lower variances than those coming from speech [2].

Numerous endpoint detection algorithms with noise immunity have been proposed [2,3,4,5]. However, some algorithms introduce additional delays as they look back a number of frames in the process of speech/non-speech classification. Our proposed solution is based on the differences exhibited by the cepstral features when computed from speech and when computed from noise and it offers the following advantages: (1) it does not require multiple passes, (2) it uses five different decision boundaries to improve classification of speech vs. silence in noisy conditions, (3) it does not set fixed thresholds, (4) it does not require any prior knowledge about the

type of noise present, and (5) it does not require the storage of pre-computed statistics since the background statistics are computed for each utterance.

2. NEWLY PROPOSED CLASSIFIER

2.1. Algorithm Description

The main idea is to characterize the background/silence at the beginning of an utterance with discriminating statistics and then relying on these statistics to classify speech vs. non-speech. The speech vs. non-speech classification relies on two measures: (1) the current frame energy, E_k , relative to the average background/silence energy level, E_{sil} , and (2) Euclidean cepstral distance, d_k , between the cepstral vector of the current frame and the mean cepstral vector, C_{mean} , representing the background/silence. Five decision boundaries are devised from these two measures to insure that low-energy sounds such as /f/, /s/, /t/, etc. are not excluded from speech, and that noise sounds are not classified as speech. A detailed description of speech vs. non-speech classification is provided in Section 2.4.

2.2. Characterization of Background Statistics

In order to characterize the background characteristics, we assume that approximately the first 100 m-sec of the utterance contains no speech. In this implementation, we assume a window size of 30 m-sec and a frame rate of 20 m-sec. The statistics of the background or silence are then based on the first N_F overlapping frames, where $N_F = 4$. The background statistics can be computed for each utterance or for each session. For file-based evaluations, the background statistics are computed for each input utterance. The background is characterized with three sets of statistics: (1) a mean silence energy value E_{sil} , (2) a mean cepstral vector C_{mean} , and (3) a mean cepstral distance D_{sil} . First, we compute the average background or silence energy, E_{sil} , by computing the energy in each of the first N_F frames from which we compute the average energy across all N_F frames. If the value of E_{sil} falls below a floor, which is a threshold value determined experimentally, then the value of E_{sil} is replaced by the floor energy value.

Next, we compute a cepstral vector, c_j , for each of the N_F frames from which we compute a mean cepstral vector, C_{mean} , according to the following:

$$C_{mean}(i) = \frac{1}{N_F} \sum_{j=1}^{N_F} c_j(i),$$

where N_F is the number of frames ($N_F = 4$), and $c_j(i)$ is the i^{th} cepstral coefficient corresponding to the j^{th} frame. The resulting vector, C_{mean} , represents the average spectral characteristics of background/silence.

Finally, we measure the Euclidean distance between each of the first N_F frames of background/silence and the mean cepstral vector, C_{mean} , representing background/silence as follows:

$$d_j = \sum_{i=1}^p (c_j(i) - C_{mean}(i))^2,$$

where d_j is the Euclidean distance between frame j and the background mean cepstral vector C_{mean} , p is the order of the cepstral analysis, $c_j(i)$ are the elements of the frame cepstral vector, and $C_{mean}(i)$ are the i^{th} elements of the background/silence mean cepstral vector, C_{mean} .

The average Euclidean distance between the first N_F frames and the average cepstral vector representing background characteristics is computed from d_j as follows:

$$D_{sil} = \frac{1}{N_F} \sum_{k=1}^{N_F} d_j.$$

If the value of D_{sil} falls below a floor, which is a pre-determined threshold value computed experimentally, then the value of D_{sil} is replaced by the floor value.

2.3. Characterization of the Input Signal

Once the statistics of the background or silence have been computed for the input utterance, then the following two parameters are computed for each frame in the input signal: (1) the frame energy E_k and (2) the Euclidean distance, d_k , between the current cepstral vector for frame k and the background mean cepstral vector, C_{mean} . The Euclidean distance, d_k , is computed according to the following equation:

$$d_k = \sum_{i=1}^p (c_k(i) - c_{mean}(i))^2,$$

where p is the order of the cepstral analysis, $c_k(i)$ are the cepstral parameters of the current frame k , and $c_{mean}(i)$ are the elements of the background mean cepstral vectors.

2.4. Description of the Classifier

Figure 1 illustrates the decision boundaries of an energy-based classifier and the decision boundaries of the proposed classifier. For the energy-based classifier, a pre-determined constant value, K , is added to the average silence energy, E_{sil} . A frame is declared as speech if its energy value exceeds $E_{sil} + K$. On the other hand, the proposed classifier determines that an input frame is speech if it satisfies any one of the following three conditions:

- (1) $d_k > \alpha * D_{sil} \quad \& \quad E_k > \beta * E_{sil}$
- (2) $d_k > D_{sil} \quad \& \quad E_k > \eta * E_{sil}$
- (3) $E_k > \gamma * E_{sil}$

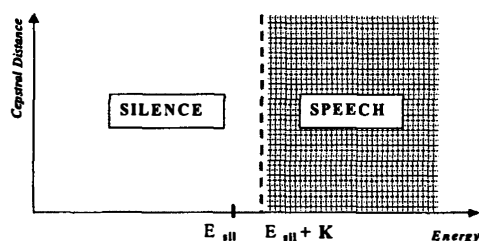
where E_{sil} is the mean silence energy, D_{sil} is the mean Euclidean cepstral distance for silence, d_k is the cepstral distance between the current frame k and the background mean cepstral vector C_{mean} , and E_k is the energy of the current frame k . The coefficients α , β , η , and γ are determined experimentally, as will be discussed later.

Condition 1 addresses the problem of low energy speech sounds in the presence of noise. In the presence of noise, the energy of a fricative is at times less than the energy threshold computed for the background statistics. In such a case, the first condition in the classifier ensures that a low-energy sound will

be classified as speech if the frame possesses similar spectral characteristics to speech, i.e., if the Euclidean spectral distance, d_k , between that frame and silence is very large.

Condition 2 addresses the problem of misclassifying sounds with high energy as speech. Clicks and pops, for example, possess energy levels that are higher than the energy threshold computed for the background statistics. To overcome this problem, condition 2 requires that in addition to having a high energy level, the input frame must also have similar spectral characteristics to speech, i.e., the Euclidean spectral distance, d_k , between that frame and silence should exceed the mean Euclidean distance for silence, D_{sil} . By relying on both energy and spectral distance, a click would not be incorrectly classified as speech.

SPEECH vs. SILENCE CLASSIFICATION BASED ON ENERGY



SPEECH vs. SILENCE CLASSIFICATION BASED ON BOTH CEPSTRAL DISTANCE & ENERGY

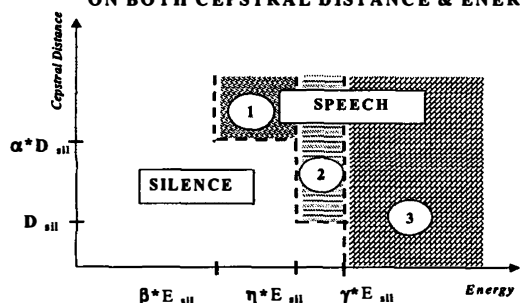


Figure 1. Speech vs. non-speech classification boundaries. Top figure shows energy-based classification decision while the lower figure shows the classification boundaries of the new algorithm ($\alpha = 3$; $\beta = 0.75$; $\eta = 1.1$; $\gamma = 1.3$).

Condition 3 relies on energy alone for the classification of an input frame. With this condition, sounds with very high energy are classified as speech regardless of their Euclidean spectral distance to the background mean cepstral vector. This condition can also help to recover from possible erroneous characterizations of the background statistics such as when the first few frames in the beginning of an utterance consist of speech rather than silence or noise. In that case, the spectral characteristics computed from the initial few frames represent speech and not silence.

The coefficients α , β , η , and γ are determined experimentally based on a subset of 3 databases: the TIDIGIT database, an in-house database collected over the telephone network as

described in [1], and an in-house database collected in-person. The statistics were based on a total of 3180 utterances. Similarly, the floor values for the background average energy, E_{sil} , and the floor value for the background Euclidean average cepstral distance, D_{sil} , are determined experimentally from the same databases.

3. EVALUATIONS

3.1. Database

Evaluations were performed for both speaker-independent and speaker dependent isolated word recognition. For the speaker-independent evaluations, each test set consists of a total of 2377 utterances representing a vocabulary size of seven command and control isolated words, each word repeated twice by 219 male and female speakers. The test data was collected over the telephone network.

For the speaker-dependent evaluations, each test set consists of 2000 utterances spoken by 10 different speakers. Each speaker spoke ten repetitions of 20 different names.

3.2. Noise Addition

For generating the noisy speech files, we follow the same procedure recommended by ITU and utilize their software and noise files. A noise segment is randomly selected from the noise file. The speech waveform and the noise waveform are each calibrated using the standardized ITU utility such that when the resulting speech and noisy waveforms are added together, we obtain the noisy speech waveform with the desired signal-to-noise ratio.

Results are presented here for 6 different types of noise conditions: car noise (urban driving at 40 mph), highway noise, office noise, babble noise, lobby noise, shopping noise at various signal-to-noise (SNR) ratios (20dB, 15dB, 10dB, 5dB).

3.3. Experimental Results

The proposed endpoint algorithm significantly improves the recognition accuracy and the false rejection rates as can be seen in Figures 2 and 3. Figure 2 illustrates the improvements in speaker-dependent recognition accuracy and in false rejection rates for 5 different noise types at 20dB SNR. Figure 3 shows the improvements in speaker-independent false rejection rates. The recognition accuracy is not explicitly shown since it is maintained. The results for SI are shown for 5 different noise types and for varying levels of SNR from 5 to 20 dB.

The improved performance is attributed to better estimates of the speech endpoints. In addition to improving recognition accuracy and false rejection rates, the proposed endpointer reduces some of the errors produced by the energy-based endpointing in the presence of noise such as (1) not finding both the beginning and ending of speech, (2) not finding the ending of speech, and (3) falsely triggering in the presence of noise and resulting in a "spoke too soon" error. The proposed endpointer reduces such errors (not plotted here) by 88% for babble noise at 20dB and 15dB SNR, and by 70% for babble noise at 10dB SNR.

4. CONCLUSIONS

We proposed a low-complexity and low-delay endpointing algorithm, which improves the performance of the recognition system, when compared to energy-based endpointing, for various SNR levels and various noise conditions including car, highway, lobby, shopping, office and babble noise. For a speaker-dependent recognizer at 20 dB SNR, it improves the recognition accuracy by 14% for lobby noise and 25% for office noise, and reduces the false rejection rates by 45% for both lobby and shopping noise. For a speaker-independent recognizer at 20 dB SNR, it improves the false rejection rates by 45% for babble noise, and 29% for both office and lobby noise. It also reduces errors due to endpointing. It is computationally efficient and introduces no further delays beyond energy-based endpointing.

5. REFERENCES

- [1] S. E. Bou-Ghazale, and A. O. Asadi, "Hands-Free Voice Activation of Personal Communication Devices," Proc. ICASSP, vol. 3, pp. 1735-1738, 2000.
- [2] J.A. Haigh, and J.S. Mason, "Robust Voice Activity Detection Using Cepstral Features," *Computer, Communication, Control and Power Engineering*. Proceedings of the IEEE Region 10 Conference TENCN, vol.3, pp. 321-324, 1993.
- [3] 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, Vol. 2, No. 3, pp. 406-412, July 1994.
- [4] L. Karray, C. Mokbel, and J. Monne, "Solutions for Robust Speech/Non-Speech Detection in Wireless Environment," Proc. IEEE 4th Workshop IVTTA'98, pp. 166-170, 1998.
- [5] D. Wu, M. Tanaka, R. Chen, L. Olorenshaw, M. Amador, and X. Menendez-Pidal, "A Robust Speech Detection Algorithm for Speech Activated Hands-Free Applications," Proc. ICASSP, vol.4, pp. 2407-2410, 1999.
- [6] X. Zhang and R. J. Mammone, "Channel and Noise Normalization Using Affine Transformed Cepstrum", Proc. ICSLP, vol. 4, pp. 1993-1996, 1996.

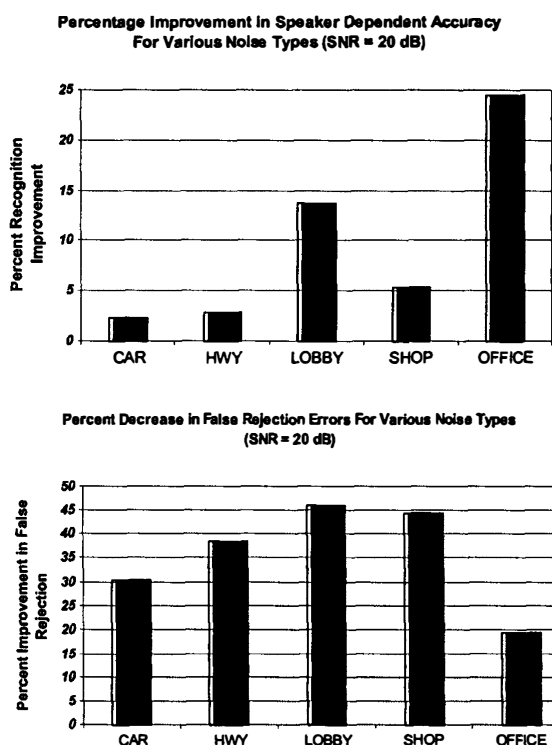


Figure 2. Improvements in speaker dependent speech recognition accuracy and false rejection rates achieved with the proposed endpoint detection as compared to energy based endpoint detection.

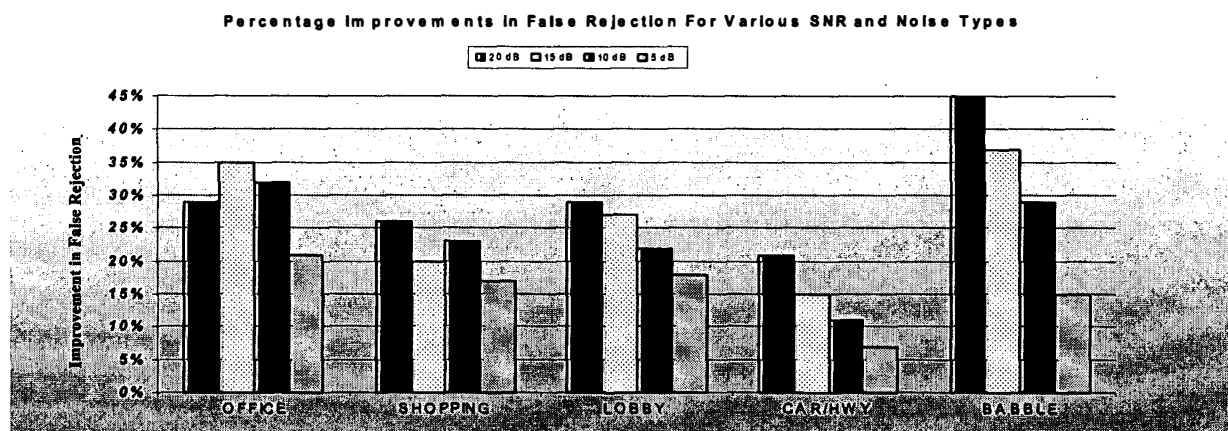


Figure 3. Improvements in SI false rejection rates achieved with the new endpoint detection as compared to energy-based endpoint detection.