ROBUST ENDPOINT DETECTION FOR SPEECH RECOGNITION BASED ON DISCRIMINATIVE FEATURE EXTRACTION

Koichi Yamamoto†, Firas Jabloun‡, Klaus Reinhard‡ and Akinori Kawamura†

†Multimedia Laboratory, Corporate R&D Center, Toshiba Corp.
‡Speech Technology Group, Cambridge Research Laboratory, Toshiba Research Europe Ltd.
†{koichi10.yamamoto, akinori.kawamura}@toshiba.co.jp, ‡{firas.jabloun, klaus.reinhard}@crl.toshiba.co.uk

ABSTRACT

Accurate endpoint detection is important for improving the speech recognition capability. This paper proposes a novel endpoint detection method which combines energy-based and likelihood ratio-based voice activity detection (VAD) criteria, where the likelihood ratio is calculated with speech/non-speech Gaussian mixture models (GMMs). Moreover, the proposed method introduces the discriminative feature extraction technique (DFE) in order to improve the speech/non-speech classification. The DFE is used in the training of parameters required for calculating the likelihood ratio. Experimental results have shown that the proposed endpointer achieves good performance compared to an energy-based endpointer in terms of start-of-speech (SOS) and end-of-speech (EOS) detections. Due to the improvement of the endpointer, the performance of automatic speech recognition (ASR) has also been improved.

1. INTRODUCTION

Robust endpoint detection is crucial for achieving good performance in automatic speech recognition (ASR) systems. In noisy conditions such as in the case of in-car applications, the endpointer fails to detect correct speech segments and causes additional errors in an ASR system. With a view to widening the usage of ASR systems in real-life environments, the demand for a robust endpointer has been growing. The goal of the work reported in this paper is to develop a robust endpointer for such an ASR system. By developing a robust endpointer, it is possible to improve ASR performance in noisy environments. Moreover, the accurate endpoint detection reduces the response time and the computation cost of ASR systems. This is because only useful speech frames are passed to a back-end decoder.

In the field of endpoint detection, energy is one of the most widely used features [1]. This is because of its simplicity and adequate performance in clean conditions. However, an energy-based method does not have robustness in low SNR conditions [2]. In order to improve the noise robustness of an energy-based endpointer, some combinations with spectrum-based features such as entropy [3] and cepstral features [4, 5] have been reported. In [4], linear discriminant analysis (LDA) is also applied to MFCC in order to extract discriminative features for speech/non-speech classification.

We propose a novel endpoint detection method which is robust even in noisy environments. The method combines the energy-based and likelihood ratio-based [6] criteria for voice activity detection (VAD), where the likelihood ratio is calculated using speech/non-speech Gaussian mixture models (GMMs). Moreover, the proposed method introduces the discriminative feature extraction technique (DFE) [7] in order to extract discriminative features for speech/non-speech classification. The DFE is used in the training of parameters required for calculating the likelihood ratio. The main advantage of introducing DFE is that DFE optimizes all the parameters of both front-end feature extractor and back-end classifier in a unified framework with a minimum classification error (MCE) criterion [8].

The rest of this paper is organized as follows. In Section 2, the conventional energy-based and likelihood-based voice activity detection (VAD) techniques are described. In Section 3, the framework of the proposed endpointer and the parameter optimization by DFE are illustrated. In Section 4, the performance of the proposed endpointer is evaluated. Finally, a conclusion is given in Section 5.

2. VOICE ACTIVITY DETECTION

2.1. Energy-based criterion

Energy is widely used as a feature for VAD. In addition to its simplicity, energy has achieved adequate performance in clean environments. In the energy-based VAD, if a log-energy exceeds a threshold, the frame is classified as speech, otherwise it is classified as non-speech. The speech threshold needs to be adjusted based on the level of the input signal. In [1, 2], adaptive threshold techniques are proposed. The noise level $E_{noise}(t)$ is estimated during non-speech segments using the following first recursive order system:

$$E_{noise}(t) = \lambda E_{noise}(t-1) + (1-\lambda)E(t)\ , \qquad (1)$$

where $E(t)$ is the log-energy of frame t and λ is the forgetting factor. The speech threshold $T_e(t)$ is then set according to the following equation:

$$T_e(t) = E_{noise}(t) + \gamma\ , \qquad (2)$$

where γ is a fixed value to determine the threshold. If $E(t) > T_e(t)$, the update in Eq. (1) stops. If $E(t) < E_{noise}(t)$, the update restarts.

2.2. Likelihood-based criterion

GMMs have been widely used as classifiers in the speaker recognition field due to their adequate modeling performance and

text-independency [9, 10]. By training one GMM with speech data and another with non-speech data, it is possible to handle the frame-based speech/non-speech classification [6]. The log-likelihood ratio of speech and non-speech GMMs is calculated as follows:

$$L(t) = g_1(\mathbf{y}(t);\Lambda) - g_0(\mathbf{y}(t);\Lambda)\,, \tag{3}$$

where g_0 and g_1 represent the log-likelihood of the non-speech and speech GMM respectively, $\mathbf{y}(t)$ represents a feature vector for frame t and Λ represents the parameter set of both speech and non-speech GMMs. These parameters are trained based on the maximum likelihood (ML) criterion with the expectation maximization (EM) algorithm. If $L(t)$ exceeds a speech threshold, the frame is classified as speech, otherwise it is classified as non-speech.

3. PROPOSED ENDPOINT DETECTION

3.1. Framework of proposed endpointer

The proposed endpointer utilizes both the energy and the likelihood ratio for the VAD. In order to improve robustness to noisy environments, spectral subtraction (SS) is used as a pre-processing step, where the noise spectrum is estimated using the quantile based noise estimation technique (QBNE) [11]. An input signal is framed using a hamming window and the power spectral density (PSD) of each frame is calculated. QBNE-SS is then applied as follows:

$$\hat{S}(k,t) = \max\{X(k,t) - \alpha\hat{N}(k,t),\, \beta X(k,t)\}\,, \tag{4}$$

where $X(k,t)$ represents the k-th PSD of the noisy signal at frame t, $N(k,t)$ represents the k-th PSD of the noise estimated by QBNE and $\hat{S}(k,t)$ represents the k-th PSD of an enhanced input signal. The parameters α and β control the subtraction and flooring value. The log-energy of the frame t is calculated by the following equation:

$$E(t) = \log \sum_{k=K_L}^{K_H} \hat{S}(k,t)\,, \tag{5}$$

where K_L and K_H represent the lowest and highest frequency components which are used to calculate the log-energy, respectively.

For the feature vector of the GMMs, a log mel-filterbank energy is utilized. In order to extract the difference of time-variation, a corresponding delta is concatenated to the log mel-filterbank energy. The first form of the feature vector $\mathbf{x}(t)$ is represented as follows:

$$\mathbf{x}(t) = \left[x_1(t), \cdots, x_N(t), \Delta_1(t) \cdots, \Delta_N(t)\right]^T\,, \tag{6}$$

where N represents the number of mel-filterbanks, $x_n(t)$ represents the n-th log mel-filterbank energy and $\Delta_n(t)$ represents the corresponding delta. The static part $x_n(t)$ of the feature vector $\mathbf{x}(t)$ changes with the level of the input signal. In order to extract only the characteristics related to the spectral shape, the feature vector $\mathbf{x}(t)$ is normalized by subtracting the mean of each frame as follows:

$$\bar{x}_n(t) = x_n(t) - m(t)\,, \tag{7}$$

where,

$$m(t) = \frac{1}{N}\sum_{n=1}^{N} x_n(t)\,. \tag{8}$$

The normalized feature vector $\bar{\mathbf{x}}(t)$ is represented as follows:

$$\bar{\mathbf{x}}(t) = \left[\bar{x}_1(t), \cdots, \bar{x}_N(t)\,, \Delta_1(t) \cdots, \Delta_N(t)\right]^T\,. \tag{9}$$

After normalization, $\bar{\mathbf{x}}(t)$ is projected to a lower feature vector $\mathbf{y}(t)$ for decorrelation and for the reduction of computational cost. The projection is represented by the following equation:

$$\mathbf{y}(t) = \mathbf{P}\bar{\mathbf{x}}(t)\,, \tag{10}$$

where $\mathbf{P}$ is an $M \times 2N$ projection matrix which is obtained using principal component analysis (PCA). After the extraction of the final form of the feature vector $\mathbf{y}(t)$, the log-likelihood ratio of speech/non-speech is calculated as in Eq. (3).

In the proposed endpointer, a frame is judged as speech only when it satisfies the following condition:

$$E(t) > T_e(t) \quad \& \quad L(t) > T_l(t)\,, \tag{11}$$

where $T_e(t)$ and $T_l(t)$ represent the speech threshold for the energy and the likelihood ratio, respectively. Both thresholds are updated adaptively based on the method described in Section 2.1. This combination makes it possible to utilize both energy and spectral information for the VAD.

After the VAD, a finite-state automaton decides the start-of-speech (SOS) and end-of-speech (EOS) points. The automaton is driven based on the frame-based classification. Some decision rules related to time constraint are used to decide both SOS and EOS.

3.2. Discriminative feature extraction

In order to calculate the likelihood ratio, it is necessary to train the parameters: the elements of the projection matrix and the means, variances, and mixture weights of the speech/non-speech GMMs. The projection matrix is obtained using PCA. The GMMs are trained by EM algorithm. However, these techniques are not based on a criterion which minimizes the speech/non-speech classification errors. Therefore, we introduce the discriminative feature extraction technique (DFE) [7] in order to optimize the parameters of both the projection matrix and the GMMs. DFE is based on the minimum classification error/generalized probabilistic descent (MCE/GPD) method [8] and can adjust a feature extractor as well as a classifier in a unified framework. It was reported to be an effective technique for GMM-based speaker recognition systems [9, 10].

In the proposed method, the frame-based misclassification measure of the likelihood ratio is defined as follows:

$$d = -g_j(\mathbf{y}(t);\Lambda) + g_{i\neq j}(\mathbf{y}(t);\Lambda)\,, \tag{12}$$

where,

$$\mathbf{y}(t) \in C_j \quad \text{and} \quad i, j \in [0,1]\,. \tag{13}$$

C_j represents the two classes (C_0 :non-speech or C_1 :speech). If the frame is classified correctly, d becomes negative. From the misclassification measure, the loss function of DFE is defined as follows:

$$l = \frac{1}{1+\exp(-\tau d)}, \tag{14}$$

where τ represents a positive parameter which controls the slope of the sigmoid function. The loss function becomes close to 1 in the case of miss-classification, otherwise it becomes close to 0. All adjustable parameters of the projection matrix and the speech/non-speech GMMs are defined as Φ. In order to minimize the loss function l in Eq. (14), the parameter set Φ is updated based on the MCE/GPD training rule:

$$\Phi[t+1] = \Phi[t] - \varepsilon_t \nabla_{\Phi} l(\bar{\mathbf{x}}(t); \Phi[t]), \tag{15}$$

where ε_t represents the step size parameter which decreases according to the number of iterations. Parameter re-estimation is applied for every frame with training data until the parameters converge.

In the adjustment process, the variances and weights of the GMMs are subject to certain constraints. They should be positive values and the summation of the weights should be one. In order to satisfy the constraints, these parameters are transformed into a parallel subspace before adjustment. The parameters are adjusted within the subspace and then transformed inversely. The details of the subspace technique are described in [9].

4. EXPERIMENTAL RESULTS

Two experiments were conducted in order to evaluate the performance of different endpointers: a conventional energy-based approach [2] enhanced by QBNE-SS and the proposed endpointer both without and with DFE training. In the first experiment, the differences between manually labeled and detected endpoints were measured [12]. The second experiment was conducted to evaluate the endpointers in terms of ASR performance.

4.1. Experimental setups

4.1.1. Training databases

For the training of the projection matrix and the GMMs, speech and noise datasets were prepared. The speech data consisted of 3000 short utterances recorded in a clean environment covering four languages: English, French, German and Japanese. The *JEIDA* noise database [13] was used as noise data. The database consisted of 18 kinds of noises: car noise, factory noise, babble noise, etc. In order to create the noisy speech data, a part of the noise data was artificially added to the speech data, where the SNRs were 0dB, 5dB, 20dB and clean.

4.1.2. Experimental conditions

An input signal was sampled at 11025Hz and framed using a hamming window. The length of one frame was 23ms with 8ms shift. The parameters K_L and K_H in Eq. (5) were set to 130Hz and 4900Hz, respectively. The number of mel-filterbanks N was set to 24 and the dimension M of the final feature vector $\mathbf{y}(t)$ was set to 16. The number of frames for extracting the delta was set to nine.

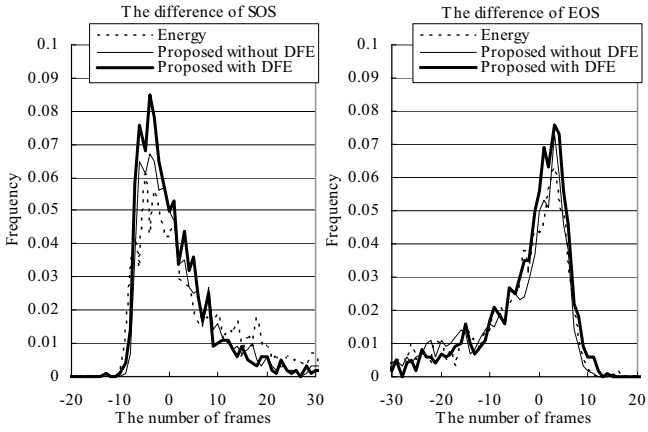


Figure 1. The histograms of the differences (# of frames) between manually labeled and detected endpoints: SOS (left) and EOS (right) points for 5dB SNR car noise.

In the DFE training, PCA and EM algorithm were used to obtain the initial values of the projection matrix and the GMMs. PCA was calculated using the 48-dimensional feature vectors as described in Eq. (9). These feature vectors were extracted from both speech and noise training data. The eigenvectors with the top-16 eigenvalues of the correlation matrix calculated from the feature vectors were chosen as the initial projection matrix, where the cumulative proportion was 0.87. As the initial classifier, 32-mixture diagonal GMMs were used. The GMMs were trained by EM algorithm, where the initial mean vectors were obtained using the LBG algorithm and the initial diagonal variances and mixture weights were set to 1 and 1/32, respectively. The DFE training was iterated 32 epochs with all speech and noise training data, where the order of the samples was set randomly for each epoch. τ in Eq. (14) was set to 1.5. ε_t in Eq. (15) was initially set to 1.0×10^{-4} and was decreased monotonically in the following epochs as the iteration increases.

4.2. Endpoint accuracy

The first experiment was evaluated in terms of differences between manually labeled and detected endpoints. The test dataset used in this experiment consisted of 1000 utterances of Japanese city names. Car noise and babble noise which were different from training data were artificially added to the database with 5dB SNR.

Figure 1 shows the histograms of the differences for the car noise and Table 1 lists the statistical information of the histograms for all conditions. In this experiment, the automatons of each endpointer were tuned using training data with a criterion which maximized the rate of a distribution less than 10-frames difference.

Table 1. The statistical information of the histograms, where each value represents the rate (%) of the distribution.

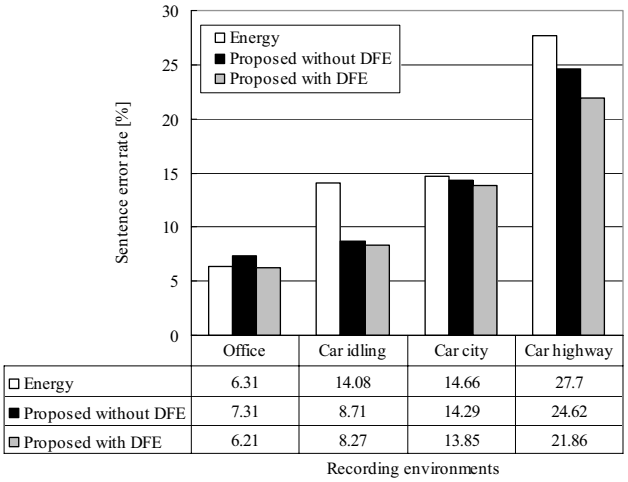
Conditions	Clean				Car 5dB				Babble 5dB			
The difference of the number of frames	SOS		EOS		SOS		EOS		SOS		EOS	
	≤ 10	≤ 30	≤ 10	≤ 30	≤ 10	≤ 30	≤ 10	≤ 30	≤ 10	≤ 30	≤ 10	≤ 30
Energy	96.7	99.7	91.7	99.1	59.5	79.7	60.3	78.4	57.1	77.0	56.9	76.3
Proposed without DFE	94.0	98.9	92.7	98.2	67.5	82.5	60.0	79.6	63.3	78.0	60.2	78.1
Proposed with DFE	95.9	99.1	92.5	98.0	79.6	92.2	73.8	90.6	79.5	91.6	74.3	91.6

In Fig. 1, the histograms of the proposed endpointers (without and with DFE) show sharper peaks than in the case of using the energy-based method. This means that the proposed endpointers achieved good performance for SOS and EOS detections. Moreover, the results obtained through DFE training outperformed the results without DFE. The differences of each endpointer can be clearly seen in Table 1 where the results for clean and babble noise are also shown. For the clean condition, all endpointers showed good performance and there is no significant difference among them. For the noisy conditions, on the other hand, the DFE training improved the endpoint accuracy of the proposed method.

4.3. Speech recognition accuracy

The second experiment was conducted in terms of ASR performance with the three endpointers. The Toshiba ASR engine is developed for an embedded platform. It uses a proprietary MFCC-based front-end and an efficient HMM-based decoder, where the total number of Gaussians is about 8000. The acoustic models are tuned for noisy in-car environments. A command and control task in English was used. The corpora for the task were recorded in four kinds of real-life environments: office, in-car idling, in-car driving in city conditions and in-car driving in highway conditions. A grammar of approximately 3700 unique utterances was used for the task, representing the total number of unique utterances in the corpora in all four environments.

Figure 2 shows the sentence error rate of the ASR for the four recording environments. The proposed endpointer without DFE outperformed the energy-based technique for in-car conditions. For the idling and highway, it achieved 38.1% and 11.1% of relative error reduction rate, respectively. The DFE training further improved the performance of the proposed endpointer for all environments. In particular, for the highway condition, it achieved 11.2% of relative error reduction rate compared to the case without DFE. These experimental results have shown that by training the parameters of the projection matrix and the GMMs with DFE, the robustness to adverse conditions is improved in terms of the ASR accuracy as well as in terms of the endpoint accuracy.



	Office	Car idling	Car city	Car highway
Energy	6.31	14.08	14.66	27.7
Proposed without DFE	7.31	8.71	14.29	24.62
Proposed with DFE	6.21	8.27	13.85	21.86

Figure 2. The sentence error rate of the ASR for the four recording environments.

5. CONCLUSION

This paper presented a robust endpoint detection method for speech recognition. The proposed endpointer is based on voice activity detection (VAD) with both energy-based and likelihood ratio-based criteria. Moreover, the proposed endpointer introduces the discriminative feature extraction technique (DFE) in order to optimize the parameters for the calculation of the log-likelihood ratio. Experimental results have shown that DFE training improves the performance of the endpointer in terms of SOS and EOS detections. In the ASR evaluation, the proposed endpointer has shown the improvement of the recognition accuracies in noisy environments compared to the conventional energy-based endpointer.

6. REFERENCES

[1] S. V. Gerven and F. Xie, "A Comparative Study of Speech Detection Methods," in *Proc. EUROSPEECH'97*, vol.III, pp.1095-1098, September 1997.

[2] P. Renevey and A. Drygajlo, "Entropy Based Voice Activity Detection in Very Noisy Conditions," in *Proc EUROSPEECH 2001*, pp.1887-1890, September 2001.

[3] L.-S. Huang, C.-H. Yang, "A Novel Approach to Robust Speech Endpoint Detection in Car Environments," in *Proc. ICASSP 2000*, vol.3, pp.1751-1754, June 2000.

[4] A. Martin, D. Charlet and M. Manuuary, "Robust Speech/Non-Speech Detection using LDA Applied to MFCC," in *Proc. ICASSP 2001*, vol.1, pp.237-240, May 2001.

[5] S. E. Bou-Ghazale and K. Assaleh, "A Robust Endpoint Detection of Speech for Noisy Environments with Application to Automatic Speech Recognition," in *Proc. ICASSP 2002*, vol.4, pp.3808-3811, May 2002.

[6] N. Binder, K. Markov, R. Gruhn and S. Nakamura, "Speech Non-Speech Separation with GMMs," in *Proc. Acoustic Society of Japan Fall Meeting*, vol.1, pp.141-142, October 2001.

[7] A. Biem, S. Katagiri and B. H. Juang, "Discriminative Feature Extraction for Speech Recognition," in *Proc. 1993 IEEE Workshop on Neural Networks for Signal Processing*, pp.392-401, September 1993.

[8] S. Katagiri, C. H. Lee and B. H. Juang, "A generalized probabilistic descent method," in *Proc. Acoustic Society of Japan Fall Meeting*, pp.141-142, September 1990.

[9] C. Miyajima, H. Watanabe, K. Tokuda, T. Kitamura and S. Katagiri, "A new approach to designing a feature extractor in speaker identification based on discriminative feature extraction," *Speech Communication*, vol.35, no.3-4, pp.203-218, October 2001.

[10] J. H. Nealand, A. B. Bradley and M. Lech, "Discriminative Feature Extraction Applied to Speaker Identification," in *Proc. ICSP'02*, pp.484-487, August 2002.

[11] V. Stahl, A. Fischer and R. Bippus, "Quantile Based Noise Estimation for Spectral Subtraction and Wiener Filtering," in *Proc. ICASSP 2000*, pp.1975-1878, June 2000.

[12] Q. Li, J. Zheng, A. Tsai and Q. Zhou, "Robust Endpoint Detection and Energy Normalization for Real-Time Speech and Speaker Recognition," *IEEE Transactions on Speech and Audio Processing*, vol.10, no.3, pp.146-157, March 2002.

[13] http://www.sunrisemusic.co.jp/dataBase/fl/noisedata01_fl.html (in Japanese)