社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2005－ＳＬＰ－59 (37)
2005／12／22

識別的特徴抽出に基づく音声区間検出の検討

山本 幸一† Firas Jabloun‡ Klaus Reinhard‡ 河村 聡典†

†東芝 研究開発センター 〒212-8582 川崎市幸区小向東芝町 1

‡Speech Technology Group, Cambridge Research Laboratory, Toshiba Research Europe Ltd.

E-mail: †{koichi10.yamamoto, akinori.kawamura}@toshiba.co.jp, ‡{firas.jabloun, klaus.reinhard}@crl.toshiba.co.uk

あらまし 本稿では，音声認識のための雑音ロバストな音声区間検出方式を提案している．提案手法は，入力信号の短時間エネルギーおよび音声/非音声 GMM を用いた尤度比の二つの基準を用いてフレーム単位の音声/非音声を判別する．このとき，尤度比を計算するためのパラメータを学習する手法として，識別的特徴抽出（DFE: Discriminative Feature Extraction）を導入している．識別的特徴抽出は，特徴抽出器と識別器を統一された枠組みで識別的に最適化する特徴を持っており，音声認識および話者認識などの分野でその効果を示している．フレーム単位の音声/非音声判別性能を評価した結果，提案手法は短時間エネルギーを基準とした手法と比較して高い性能を示した．また，提案手法を用いることにより，音声の始終端検出精度および雑音環境における音声認識性能も改善された．

キーワード 音声区間検出，VAD，識別的特徴抽出，DFE，GMM

A Study on 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.

E-mail: †{koichi10.yamamoto, akinori.kawamura}@toshiba.co.jp, ‡{firas.jabloun, klaus.reinhard}@crl.toshiba.co.uk

Abstract Accurate endpoint detection is important to improve 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 method (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. Our experimental evaluation showed that the proposed method reduces the recognition error rate compared to a conventional energy-based technique.

Keyword Endpoint detection, VAD, DFE, GMM

1. INTRODUCTION

Endpoint (start- and end-of-speech) detection, which is a method to detect speech segments from input signals, is required in various speech applications such as automatic speech recognition (ASR) and speech coding. However, conventional endpoint detection methods are not robust in noisy places such as car cabins. One of the main reasons for the poor robustness is the inability to detect the endpoints in noisy places. In order to expand the speech applications even to noisy places where people carry out their daily activities, it is important to realize the robust endpoint detection. 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 endpoint detection, energy level observation is a widely used method [1]. It is popular because the method

is simple and performs well as long as environments are quiet. However, the energy-based method is not robust in low SNR environments [2]. In order to improve the robustness of energy-based endpointer which is the function to detect the start-of-speech (SOS) and the end-of-speech (EOS), some methods are reported. They are the combinations of spectrum-based methods such as entropy [3] and cepstral methods [4, 5]. Also in [4], linear discriminant analysis (LDA) is applied to the mel-frequency cepstrum coefficient (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 places. The method combines the energy-based and likelihood ratio-based [6] criteria for Voice Activity Detection (VAD). Moreover, the proposed method introduces the discriminative feature extraction method (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 the DFE introduction is that DFE optimizes all the parameters of both the front-end feature extractor and the back-end classifier in a unified framework with a minimum classification error (MCE) criterion [8].

The rest of this paper is organized as follows. Section 2 describes the conventional energy-based and likelihood-based voice activity detection (VAD) methods. In Section 3, the framework of the proposed endpointer and the parameter optimization by the DFE are described. Section 4 shows our experimental evaluations. Finally, conclusion is in Section 5.

2. VOICE ACTIVITY DETECTION

2.1. Energy-based criterion

The energy is widely used as a feature for the VAD. In addition to its simplicity, the 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:

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

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:

$$T_e(t) = E_{noise}(t) + \gamma_e, \quad (2)$$

where γ_e 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

The GMMs have been widely used as classifiers in various fields such as speaker recognition [9, 10] and audio classification [11]. By training one GMM with speech data and another GMM 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 are calculated as follows:

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

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 estimate (MLE) 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

3.1.1 Energy calculation

The proposed endpointer utilizes both the energy-based and the likelihood ratio-based criteria for the VAD. In order to improve robustness to noisy environments, the spectral subtraction (SS) is used as a pre-processing step. The noise spectrum is estimated using the quantile based noise estimation (QBNE) technique [12], where the median quantile of each PSD component within pre-determined time window is regarded as an estimated noise component. The QBNE does not need the information of voice activity, therefore it is suitable to the noise estimator for the endpointer.

An input signal is framed using a hamming window and the PSD of each frame is calculated. QBNE-SS is then applied as follows:

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

where $X(k,t)$ represents the k-th PSD of the noisy signal at frame t, $\hat{N}(k,t)$ represents the k-th PSD of the noise estimated by the 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:

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

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

3.1.2 Likelihood ratio calculation

For the feature vector of the GMMs, a log mel-filterbank energy is utilized. In order to extract the difference of time-variation like [13], 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:

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

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. 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:

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

where,

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

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

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

The normalization is applied after calculating the delta for each frame. After the 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:

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

where $\mathbf{P}$ is an $M \times 2N$ projection matrix which is obtained using the 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).

3.1.3 Finite-state automaton

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

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

where $T_e(t)$ and $T_l(t)$ represent the speech threshold for the energy and the likelihood ratio, respectively. This combination makes it possible to utilize both energy and spectral information for the VAD. As for the threshold of the likelihood ratio, it can be fixed to pre-determined value. In our preliminary experiments, however, the adaptive threshold showed better performance for the speech/non-speech classification especially in noisy conditions compared to fixed one. Therefore, in this paper, both thresholds are updated adaptively based on the method described in Section 2.1.

After the VAD, a finite-state automaton [4] 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 the PCA. The GMMs are trained by the EM algorithm. These techniques are not based on a criterion which minimizes the speech/non-speech classification errors. Therefore, we introduce the discriminative feature extraction method (DFE) [7] in order to optimize the parameters of both the projection matrix and the GMMs. The DFE is based on the minimum classification error/generalized probabilistic descent (MCE/GPD) method [8] and adjusts a feature extractor as well as a classifier in a unified framework. It was reported as an effective technique for GMM-based speaker recognition systems [9, 10].

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

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

where,

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

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:

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

where α 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:

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

where ε_t represents the step size parameter which decreases as the number of iterations increases. 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. 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

Three experiments were conducted in order to evaluate the performance of different endpointers: a conventional energy-based approach [2] enhanced by the QBNE-SS and the proposed endpointer both without and with DFE training. In the first experiment, the frame-based speech/non-speech classification was measured. In the second experiment, the differences between manually labeled and detected endpoints were measured. The final 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 [14] was used as noise data. The database consisted of 18 kinds of noises: car noise, factory noise, babble noise, etc. 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 is set to 9.

In the DFE training, the PCA and the EM algorithm were used to obtain the initial values of the projection matrix and the GMMs. The 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 the 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 decided randomly for each epoch. τ in Eq. (14) was set to 1.5. ε_l in Eq. (15) was initially set to 1.0×10^{-4} and was decreased monotonically in the following epochs.

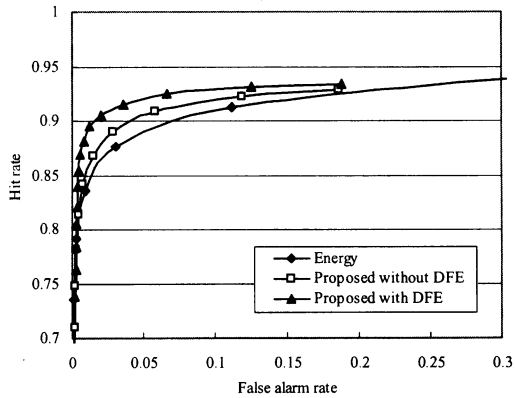

Figure 1. ROC curves for 5dB SNR car noise.

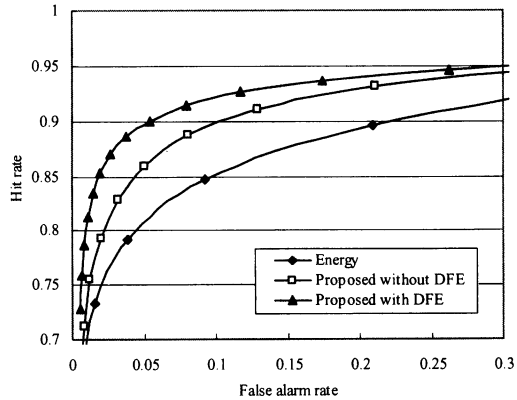

Figure 2. ROC curves for 5dB SNR babble noise.

4.2. VAD accuracy

The first experiment was evaluated in terms of frame-based speech/non-speech classification. The test dataset used in this experiment consisted of 1000 utterances of Japanese city names. The car noise and babble noise which were different from training data were artificially added to the database with 5dB SNR.

Figure 1 and 2 show the receiver operating characteristic (ROC) curves [15] for car noise and babble noise, respectively. The parameter γ_e in Eq. (2) for the energy-based technique was changed from 0.2 to 2.0 with a step size of 0.2. In the case of the proposed techniques, γ_e was set to the optimal value 0.5 and γ_l for the likelihood ratio-based criterion was changed from 0.2 to 3.0 with a step size of 0.2.

Both figures clearly show that the proposed techniques

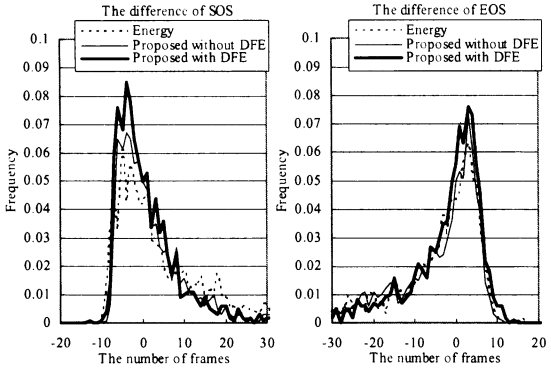

Figure 3. The histograms of the differences (# of frames) between manually labeled and detected endpoints: SOS (left) and EOS (right) points for 5dB SNR car noise.

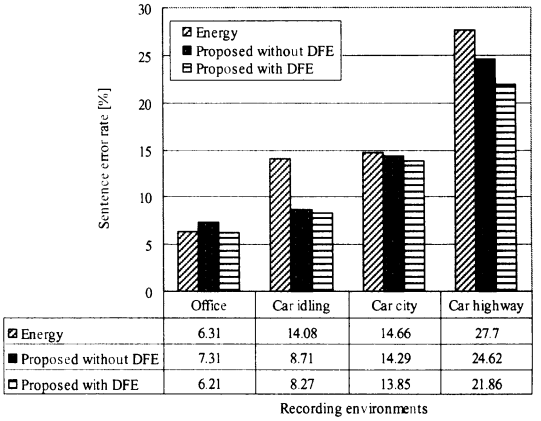

Figure 4. The sentence error rate of the ASR for the four recording environments.

(without and with DFE) achieve good speech/non-speech classification performance compared to the energy-based technique. Moreover, the results obtained through DFE training outperformed the results without DFE.

4.3. Endpoint accuracy

The second experiment was evaluated in terms of differences between manually labeled and detected endpoints. The test dataset used in Section 4.2 was evaluated.

Figure 3 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 to maximize the rate of a distribution less than 10-frames difference for training data. In Fig. 3, the histograms of the proposed endpointers (without and with DFE) show sharper peaks compared to the energy-based technique. This means that the proposed endpointers achieve good performance for SOS and EOS detections. The proposed technique with DFE training outperformed without DFE. The differences of each endpointer are 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 technique.

4.4. Speech recognition accuracy

The third 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 number of states is 800 and the number of Gaussians per state is 10. 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 actual 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 4 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

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

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 VAD and endpoint accuracies.

5. CONCLUSION

This paper presented a robust endpoint detection technique 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 discriminative feature extraction method (DFE) in order to train 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 as well as the frame-based speech/non-speech classification. 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.1883-1886, 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] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. ICASSP 1997*, pp.1331-1334, April 1997.

[12] 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.

[13] J. Padrell, D. Macho and C. Nadeu, "Robust Speech Activity Detection Using LDA Applied to FF Parameters," in *Proc. ICASSP 2005*, pp.557-560, March 2005.

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

[15] O.-W. Kwon and E.-W. Lee, "Optimizing Speech/Non-Speech Classifier Design Using AdaBoost," in *Proc. ICASSP 2003*, pp.436-439, April 2003.