Voice Activity Detection based on Optimally Weighted Combination of Multiple Features

Yusuke Kida and Tatsuya Kawahara

School of Informatics, Kyoto University,
Sakyo-ku, Kyoto 606-8501, Japan

Abstract

This paper presents a voice activity detection (VAD) scheme that is robust against noise, based on an optimally weighted combination of features. The scheme uses a weighted combination of four conventional VAD features: amplitude level, zero crossing rate, spectral information, and Gaussian mixture model likelihood. This combination in effect selects the optimal method depending on the noise condition. The weights for the combination are updated using minimum classification error (MCE) training. An experimental evaluation under three types of noisy environment demonstrated the noise robustness of our proposed method. Adapting the feature weights was shown to enhance the detection ability and to be possible using ten or fewer training utterances.

1. Introduction

One of the most significant and tackled problems in automatic speech recognition is achieving robustness against noise. The approaches to this problem include noise reduction, such as spectral subtraction and Wiener filtering, and adaptation of acoustic model to a noisy environment by MLLR or PMC. Voice activity detection (VAD) is another crucial part of the effective performance of an automatic speech recognition system [1]. If speech segments are not correctly detected, the subsequent recognition processes would be meaningless. Many methods have been proposed so far, but no single method performs satisfactorily. In addition, these VAD methods are affected by noise conditions.

To develop a VAD scheme that is robust against various kinds of noise, we propose a combination of multiple features. Our aim is to broaden the coverage of noise conditions compared to conventional VAD methods. A similar approach to noise reduction was reported in [2]. We use a combination of the following four representative features for our VAD:

- Amplitude level
- Zero crossing rate (ZCR)
- Spectral information
- Gaussian mixture model (GMM) likelihood

These features are weighted and combined, where feature weights are optimized based on minimum classification error (MCE) training. This combination in effect selects the optimal method based on the noise condition. The optimal combination of features is expected to lead to further improvement in detection accuracy. In this scheme, the data necessary to optimize the feature weights is important. We assume ten utterances are available to determine the optimal weights.

This paper is organized as follows. In Section 2, we present an overview of our VAD system scheme, each method used in the combination, we then describe the MCE training used to optimize the combination weights. The experimental conditions and results are reported in Section 3, and conclusions and future works are given in Section 4.

2. Weighted Combination of VAD Methods

2.1. Framework

The flow of our VAD system is shown in Figure 1. Our framework is applied in a framewise manner. At first, the system divides input data into frames. Then four features are calculated: amplitude level, ZCR, spectral information, and GMM likelihood. The features are shown as $f^{(1)}, \cdots, f^{(4)}$ in the figure, and they are combined with weights $w_1, \cdots, w_4$. The combined score of data frame $\boldsymbol{x}_t$ (t: frame number) is defined as follows:

$$F(\boldsymbol{x}_t) = \sum_{k=1}^{4} w_k \cdot f^{(k)}(\boldsymbol{x}_t), \tag{1}$$

where K denotes the number of combined features. The weights w_k must satisfsy the following conditions:

$$\sum_{k=1}^{K} w_k = 1, \tag{2}$$

$$w_k > 0, \tag{3}$$

where the initial weights are all equal (i.e., 0.25 in this case).

The following two discriminative functions judge whether each frame is speech or noise.

$$g_s(\boldsymbol{x}) = F(\boldsymbol{x}_t) - \theta, \tag{4}$$

$$g_n(\boldsymbol{x}) = \theta - F(\boldsymbol{x}_t), \tag{5}$$

where θ denotes the threshold value of the combined score. Data $\boldsymbol{x}_t$ is regarded as a speech frame if the discriminative function of speech $g_s(\boldsymbol{x}_t)$ is larger than that of noise $g_n(\boldsymbol{x}_t)$. Otherwise, $\boldsymbol{x}_t$ is regarded as a noise frame. Actually, this judgement can be made simply by comparing $F(\boldsymbol{x}_t)$ and θ. However, MCE training requires a discriminative function for each cluster. Therefore, the two functions are prepared. A label file is used to train the weight of each feature. This file includes hand-labeled starting and end points of each utterance. Each frame of the utterance is judged as speech or non-speech, and accordingly, the weights are updated in a framewise manner.

2.2. Features and Methods for VAD

2.2.1. Amplitude level

Amplitude level is one of the most common features of VAD methods and is used in many applications. The amplitude level

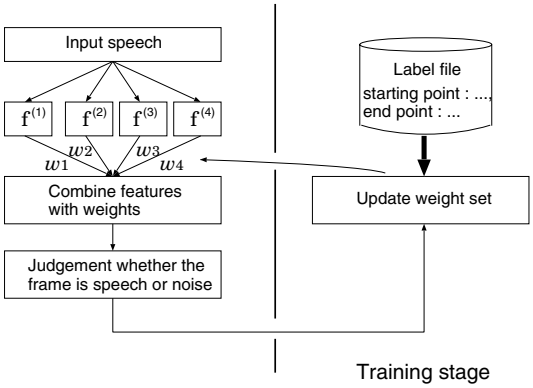

Figure 1: Framework

at the t-th frame E_t is computed as the logarithm of the signal energy; that is, for N-length Hamming-windowed speech samples $\{s_n, n = 1, \cdots, N\}$,

$$E_t = \log \sum_{n=1}^{N} s_n^2. \tag{6}$$

In our method, the amplitude level of noise is assumed to be known in advance. Then, the feature used in the combination is calculated using the ratio of amplitude of the input frame to the amplitude of noise as follows:

$$f_t^{(1)} = \frac{E_t}{E_n}, \tag{7}$$

where E_n denotes the amplitude level of noise.

2.2.2. Zero crossing rate (ZCR)

Zero crossing rate (ZCR) is the number of times the signal level crosses 0 during a fixed period of time, and it is used for not only speech but also various detection applications. Similarly to amplitude level, a ratio of the input frame to noise is used for this feature. The feature $f_t^{(2)}$ is calculated as follows:

$$f_t^{(2)} = \frac{Z_t}{Z_n}, \tag{8}$$

where Z_t denotes the ZCR of the input frame, and Z_n denotes that of noise.

2.2.3. Spectral information

Many VAD methods based on spectral information have been studied recently [4, 5]. The spectrums of speech and noise are shown in Figure 2. As shown in the figure, we partition the frequency domain into several channels and calculate the signal to noise ratio (SNR) for each channel. We then compute the average value of each SNR. The spectral information feature $f_t^{(3)}$ is defined as

$$f_t^{(3)} = \frac{1}{B} \sum_{b=1}^{B} 10 \log_{10} \frac{S_{bt}^2}{N_b^2}, \tag{9}$$

where B denotes the number of channels. The term S_{bt} and N_b indicate the average intensity within channel b for speech and noise. Similarly to amplitude level and ZCR, N_b is assumed to be obtained in advance.

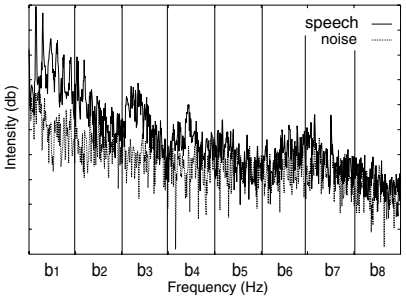

Figure 2: Spectrum of speech and noise

2.2.4. GMM likelihood

Gaussian mixture model (GMM) is getting widely used for speech detection, because of its text-independency and scalability in training [3]. A log-likelihood ratio of speech GMM to noise GMM for input frames is used for the GMM feature. The feature $f_t^{(4)}$ is calculated as

$$f_t^{(4)} = \log(p(\boldsymbol{x}_t|\Theta_s)) - \log(p(\boldsymbol{x}_t|\Theta_n)), \tag{10}$$

where Θ_s and Θ_n denote the model parameter set of GMM for the speech and noise, respectively.

2.3. Weight Optimization Using MCE Training

To adapt our VAD scheme to noisy environments, we applied MCE training based on the generalized probabilistic descent method [6] to the optimization of the weights.

2.3.1. Definition of Loss Function

For the MCE training, the misclassification measure of training data frame $\boldsymbol{x}_t$ is defined as

$$d_k(\boldsymbol{x}_t) = -g_k(\boldsymbol{x}_t) + g_m(\boldsymbol{x}_t), \tag{11}$$

where k denotes the true cluster (i.e., speech (s) or noise (n)), and m indicates another cluster. When (11) is negative $\boldsymbol{x}_t$ is correctly classified.

The loss function l_k is defined as a differential sigmoid function approximating the 0-1 step loss function:

$$l_k(\boldsymbol{x}_t) = \big(1 + \exp(-\gamma \cdot d_k)\big)^{-1}, \tag{12}$$

where γ denotes the gradient of the sigmoid function. The goal of the discriminative training is to minimize the loss function based on the probabilistic descent method.

2.3.2. Weight Adjustment

During the weight adjustment in the MCE training, the weight set $\boldsymbol{w}$ is transformed into a new set $\tilde{\boldsymbol{w}}$ because of a constraint ($w_k > 0$);

$$\tilde{\boldsymbol{w}} = \{\tilde{w}_1, \tilde{w}_2, \ldots, \tilde{w}_K\}, \tag{13}$$

$$\tilde{w}_k = \log w_k. \tag{14}$$

The weight set, $\tilde{\boldsymbol{w}}$, is sequentially adjusted every time a frame is given (i.e., sample-by-sample mode). The weight adjustment is defined as:

$$\tilde{\boldsymbol{w}}(t+1) = \tilde{\boldsymbol{w}}(t) - \varepsilon_t \nabla l_k(\boldsymbol{x}_t), \tag{15}$$

where ε_t is a monotonically decreasing learning step size. The gradient of Eq. (15) is obtained as follows.

$$\nabla_{\tilde{\boldsymbol{w}}} l_k(\boldsymbol{x}_t) = \frac{\partial l_k}{\partial d_k} \frac{\partial d_k}{\partial g_j} \cdot \nabla_{\tilde{\boldsymbol{w}}} g_j(\boldsymbol{x}_t), \tag{16}$$

where $\frac{\partial l_k}{\partial d_k}$, $\frac{\partial d_k}{\partial g_j}$, and $\nabla_{\tilde{w}_l} g_j(\boldsymbol{x}_t)$, which are elements of $\nabla_{\tilde{\boldsymbol{w}}} g_j(\boldsymbol{x}_t)$, are given by

$$\frac{\partial l_k}{\partial d_k} = \gamma \cdot l_k(1 - l_k), \tag{17}$$

$$\frac{\partial d_k}{\partial g_j} = \begin{cases} -1 & j = k \\ 1 & j \neq k \end{cases}, \tag{18}$$

$$\begin{aligned} \nabla_{\tilde{w}_l} g_j(\boldsymbol{x}_t) &= \frac{\partial}{\partial w_l} \left[\sum_{l=1}^{L} w_l F_l(\boldsymbol{x}_t) \right] \cdot \frac{\partial w_l}{\partial \tilde{w}_l} \\ &= w_l F_l(\boldsymbol{x}_t). \end{aligned} \tag{19}$$

After $\tilde{w}$ is updated, $\tilde{w}$ is returned to w as follows:

$$w_k = \frac{\exp \tilde{w}_k}{\sum_{l=1}^{L} \exp \tilde{w}_l} \tag{20}$$

Eq. (20) includes normalization of the weights, which satisfies the condition (2).

3. Experimental Evaluation

3.1. Task and Conditions

We conducted speech detection experiments in noisy environments to evaluate the performance of our proposed method. The frame-based false alarm rate (FAR) and false rejection rate (FRR) were used as evaluation measures. FAR is the percentage of non-speech frames incorrectly classified as speech, and FFR is the percentage of speech frames incorrectly classified as non-speech.

In our experiments, speech data (16kHz) from ten speakers were used. Ten utterances were used for testing for each speaker. Each utterance lasted a few seconds, and three-seconds pauses were inserted between them. To make the noisy data, we added the noises of sensor room, machine, and background speech to the clean speech data by varying SNR (10, 15dB). In total, we had 600 (= 3 types $\times$ 2 SNR $\times$ 10 persons $\times$ 10 utterances) samples as the test set. A different set of ten utterances, whose text is different from the training set, was used for the weight training for each condition.

The frame length was 100-ms for amplitude level and ZCR, and 250-ms for spectral information, and GMM likelihood. The frame shift was 250-ms for each feature. Noise features such as E_n in Eq. (7), Z_n in Eq. (8) and N_b^2 in Eq. (9) were calculated using the first second of speech data, which did not include utterances.

For GMM likelihood, a 32-component GMM with diagonal covariance matrices was used to model speech and noise. The GMM parameters were trained with EM algorithm. The parameters of GMM were 12 mel-cepstral coefficients with their Δ and Δ-power. JNAS (Japanese Newspaper Article Sentences) corpus that includes 306 people and about 32000 utterances was used to train the speech GMM. For the noise GMM, three types of noise of sensor room, office, and corridor noise were used for training. Office and corridor noise was not used to make the training and testing data for the VAD experiment.

3.2. Results

The experimental results for six types of noise conditions are shown in Figure 3∼8. These figures compare our proposed method to the individual methods we combined. The horizontal axis corresponds to the FAR, and the vertical axis corresponds to the FRR. 'Amplitude' indicates the amplitude level, 'ZCR' the zero crossing rate, 'Spectrum' the spectral information, 'GMM' the GMM likelihood, and 'Proposed' our proposed method. The operating curve is plotted by varying the threshold (θ) value of the evaluation function and each point in the figures indicates one threshold value. Under sensor room noise, 'ZCR' had the best performance of all the individual methods. Notice that GMM covered this noise in training, but did not have the best performance. For craft machine noise, 'Spectrum' performed best, and for background speech noise, 'GMM' performed the best. These observations indicate that the best method differs depending on the noise condition. However, 'Proposed' outperformed the individual methods under all noise conditions. This proves that our method is robust against noise. We also conducted a closed test, where testing was done with the speech data used for the weight training, and the result is shown as 'Closed' in Figure 3. The difference between the 'Closed' and 'Proposed' results is trivial. Thus, the proposed method is robust against variation of utterances. The same conclusion was obtained under the other conditions.

We also compared our method before and after optimizing the weights to confirm the effectiveness of the training. Before the training, the weights are set to equal (= 0.25). The equal error rate (EER) under each noise type where the SNR was 10dB is shown in Table 5. EER is a figure at the operating point where the FAR equals to the FRR. For craft machine and background speech noise, EER was significantly improved when the weights were optimized, though it was not slightly degraded under sensor room noise.

4. Conclusions

This paper proposed a VAD scheme that is robust against noise based on optimally weighted combinations of multiple features. We conducted a detection experiment under three different types of noise and compared our proposed method to the individual methods we combined. Our method performs better than any of its component methods under all noise conditions, and we found that it is robust against noise. We also confirmed that adjusting the feature weights enhances the detection ability and that our VAD system is adapted to the noise condition using only ten or fewer utterances.

Our future work includes integration of our VAD method into automatic speech recognition systems and evaluation with recognition accuracy.

5. Acknowledgements

The authors are grateful to Dr. Chiyomi Miyajima at Nagoya University for her helpful advice.

6. References

[1] R.Nishimura, Y.Nishihara, R.Tsurumi, A.Lee, H.saruwatari, K.Shikano,"Takemaru-kun: Speech-oriented information system for real world research platform,"*Int'l Workshop on Language Understanding and Agents for Real World Interaction*, pp.70-78, 2003.

[2] T.Yamada, J.Okada, K.Takeda, N.Kitaoka, M.Fujimoto,

Table 1: EER before and after weight training

noise type (10db)	before (%)	after (%)
sensor room	4.6	4.9
machine	8.9	7.3
background speech	11.2	10.8
average	8.2	7.6

S.Kuroiwa, K.Yamamoto, T.Nishiura, M.Mizumachi, S.Nakamura, "Integration of noise reduction algorithms for Aurora2 task," *EUROSPEECH-2003*, pp.1769-1772, Sep.2003.

[3] A.Lee, K.Nakamura, R.Nishimura, H.Saruwatari, K.Shikano, "Noise Robust Real World Spoken Dialog System using GMM Based Rejection of Unintended Inputs,"*ICSLP-2004*, Vol.I, pp.173-176, Oct.2004.

[4] J.Ramirez, J.C.Segura, C.Benitez, A.de la Torre, A.Rubio, "Voice Activity Detection with Noise Reduction and Long-Term Spectral Divergence Estimation,"*ICASSP-2004*, Vol.II, pp.1093-1096, May.2004.

[5] P.N.Garner, T.Fukada, Y.Komori, "A Differential Spectral Voice Activity Detector,"*ICASSP-2004*,Vol.I, pp.597-600, May.2004.

[6] B.-H.Juang, S.Katagiri, "Discriminative learning for minimum error classification," *IEEE Trans. Signal Process.*, vol.40, no.12, pp.3043–3054, Dec. 1992.

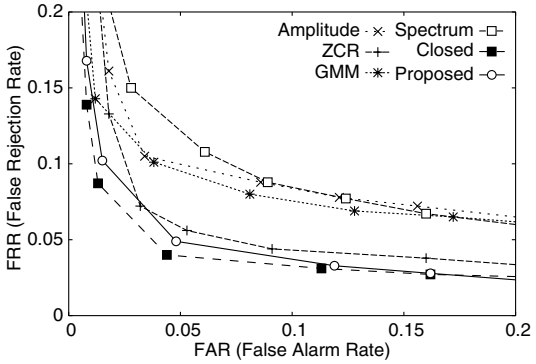

Figure 3: Sensor room:10db

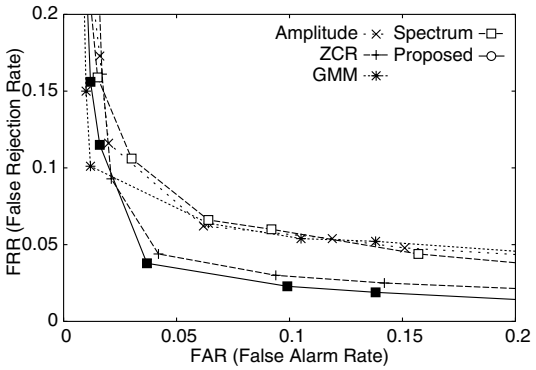

Figure 4: Sensor room:15db

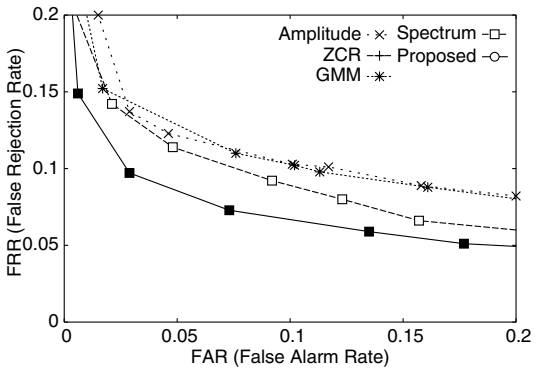

Figure 5: Craft machine:10db (Plot of ZCR is out of this range.)

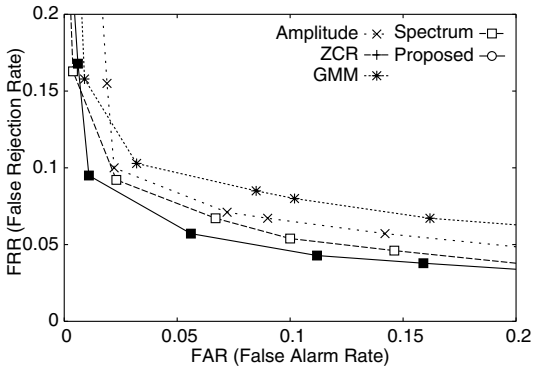

Figure 6: Craft machine:15db (Plot of ZCR is out of this range.)

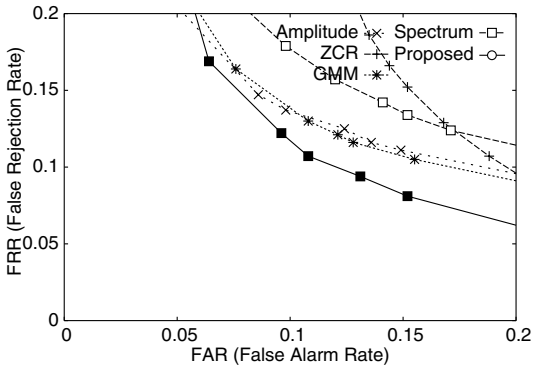

Figure 7: Background speech:10db

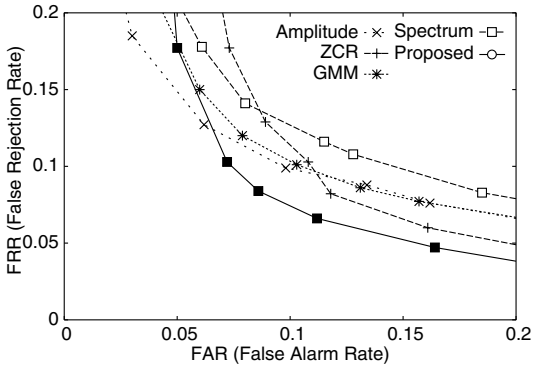

Figure 8: Background speech:15db