Evaluation of Voice Activity Detection by Combining Multiple Features with Weight Adaptation

Yusuke Kida and Tatsuya Kawahara

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

Abstract

For noise-robust automatic speech recognition (ASR), we propose a novel voice activity detection (VAD) method based on a combination of multiple features. The scheme uses a weighted combination of four conventional VAD features: amplitude level, zero crossing rate, spectral information, and Gaussian mixture model (GMM) likelihood. The weights for combination are adaptively updated using minimum classification error (MCE) training. In this paper, we first investigate the effect of adaptation of the combination weights and GMM parameters, and demonstrate that the weights can be effectively adapted with a single utterance. Then, we present application of the method to ASR. It is confirmed that the proposed method significantly outperforms conventional methods in various noise conditions.

Index Terms: speech recognition, voice activity detection, MCE training, noise adaptation

1. Introduction

Voice activity detection (VAD) is a vital front-end in automatic speech recognition (ASR) systems, especially to perform robustly in noisy environments. If speech segments are not correctly detected, the subsequent recognition processes would be often meaningless. Therefore, many studies have been conducted so far[1]-[6]. However, there are a variety of noise conditions and no single method is expected to cope with all of them.

In order to realize VAD robust against various kinds of noise, we have proposed a combination of multiple features[7]. The goal of the scheme is to broaden the coverage of noise conditions. It is somewhat similar to multi-condition training, which is very popular in acoustic modeling for noisy speech recognition, and more recently a noise reduction method integrating several techniques was proposed as well[8].

Specifically, we adopt a combination of four representative features for our VAD: amplitude level, zero crossing rate (ZCR), spectral feature, and Gaussian mixture model (GMM) likelihood. These features are combined with weights, which are optimized based on minimum classification error (MCE) training. This combination in effect selects the optimal method according to the noise condition, and the optimization of feature weights is regarded as "adaptation" to the environment. In real-world applications, it is necessary that this adaptation is made possible with a very few number of utterances. In this paper, therefore, we present investigation of the weight optimization as well as GMM adaptation. Based on the analysis, we apply the VAD method to an ASR system, to demonstrate that the enhanced VAD actually leads to improvement of ASR performance.

This paper is organized as follows. Section 2 gives a brief overview of our VAD scheme and description of adopted features. Section 3 reports experimental evaluation of the VAD, focusing on the adaptation of the combination weights and GMM parameters. Section 4 addresses its application to ASR and experimental evaluation in various noise conditions. Section 5 concludes the paper.

2. Proposed VAD Method

2.1. Framework

The flow of the proposed VAD scheme is shown in Figure 1. At first, input data is divided into frames. Then, for each frame, four features are calculated: amplitude level, ZCR, spectral information, and GMM likelihood. The features are denoted as $x^{(1)}, \cdots, x^{(4)}$ in the figure, and they are normalized through a sigmoid function as below.

$$f^{(i)}(x^{(i)}) = \frac{1}{1+\exp\{-\alpha^{(i)}(x^{(i)}-\beta^{(i)})\}} \quad (1)$$

Here, $\alpha^{(i)}$ and $\beta^{(i)}$ are determined from the variance and the mean of $x^{(i)}$. Then, the normalized features are combined with weights $w_1, \cdots, w_4$. Thus, the combined score for a frame $\boldsymbol{X}_t$ is defined as

$$F(\boldsymbol{X}_t) = \sum_{i=1}^{4} w_i \cdot f^{(i)}(x_t^{(i)}). \quad (2)$$

Here, the weights w_i must satisfy the following conditions:

$$\sum_{i=1}^{4} w_i = 1, \quad w_i > 0.$$

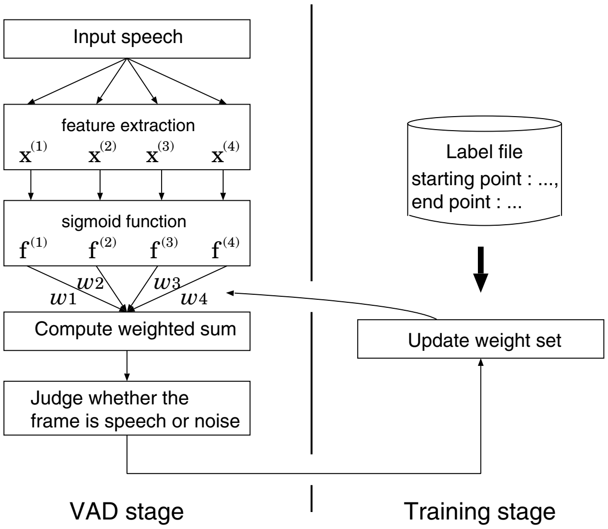

Figure 1: Framework of the proposed VAD

The initial weights are set to all equal (i.e. 0.25). The frame-wise decision of speech and non-speech is made by comparing the function $F(\boldsymbol{X}_t)$ against a threshold θ.

In the training or adaptation phase, correct labels, i.e. starting and end points of utterances, are given to compute a loss function based on $F(\boldsymbol{X}_t)$, and then MCE training is performed frame by frame.

2.2. Features for VAD

2.2.1. Amplitude level

Amplitude level is one of the orthodox features for VAD, though it is not robust against low SNR conditions[1]. It is defined as the logarithm of the signal energy; that is, for N-length Hamming-windowed speech samples $\{s_n : n = 1, \cdots, N\}$, it is computed as

$$x_t^{(1)} = \log \sum_{n=1}^{N} s_n^2. \qquad (3)$$

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 also widely used for VAD[2]. It is very effective for some kinds of noise, but not at all for noise having frequent zero crossing.

2.2.3. Spectral information

Recently, many VAD methods based on spectral information have been studied [3, 4]. We partition the frequency domain, computed with FFT, into several channels and calculate the signal to noise ratio (SNR) for each channel. For the spectral feature in our method, we compute the average SNR over all channels by

$$x_t^{(3)} = \frac{1}{B} \sum_{b=1}^{B} 10 \log_{10} \frac{S_{bt}^2}{N_b^2}, \qquad (4)$$

where B denotes the number of channels (20 in this work). The term S_{bt} and N_b indicate the average intensity within a channel b for speech and noise. Here, N_b is estimated in advance using the beginning segment of the utterance.

2.2.4. GMM likelihood

Gaussian mixture model (GMM) is getting widely used for speech detection, because the statistical model is easily trained and usually powerful[5, 6]. The log-likelihood ratio of speech GMM to noise GMM for an input frame is computed by

$$x_t^{(4)} = \log(p(v_t|\Theta_s)) - \log(p(v_t|\Theta_n)), \qquad (5)$$

where v_t is an acoustic vector for the GMMs, and Θ_s and Θ_n denote the model parameter set for speech and noise, respectively. The GMM for noise has to cover a variety of noise characteristics and should be adapted to the environment if possible.

2.3. Weight Optimization using MCE Training

To adapt our VAD scheme to noisy environments, we apply minimun classification error (MCE) training based on generalized probabilistic descent (GPD) to optimization of the combination weights w_i. Loss functions for speech and non-speech are defined, and a formula for updating weights is derived from them. Detail of the procedure is described in [7].

3. Evaluation of Weight Adaptation in VAD

3.1. Task and Conditions

We first conducted experimental evaluation of the VAD in speech detection performance in various noisy environments. The frame-wise 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 FRR is the percentage of speech frames incorrectly classified as non-speech. In this paper, we mainly use equal error rate (EER), a figure at the operating point where FAR equals to FRR, for simplicity.

In the experiments, speech data (16kHz sampling) from ten speakers were used. Ten utterances were used for testing for each speaker. Each utterance lasted a few seconds, and three-second pauses were inserted between them. To make a test set of noisy data, we added the noise of air conditioner (AC), craft machine (CM) and background speech (BS) to

Table 1: Comparison of methods (EER(%) in VAD): 5db

	GMM only		Proposed	
	(base)	(adapted)	(base)	(adapted)
AC	16.6	16.1	10.5	10.4
CM	16.8	16.4	14.4	14.3
BS	18.0	17.5	17.9	17.8

Table 2: Comparison of methods (EER(%) in VAD): 15db

	GMM only		Proposed	
	(base)	(adapted)	(base)	(adapted)
AC	7.8	7.3	4.1	4.2
CM	11.2	11.1	7.7	7.8
BS	10.3	10.4	10.0	10.1

(base): baseline GMM
(adapted): after MLLR adaptation of GMM
AC: air conditioner, CM: craft machine, BS: background speech

the clean speech by varying the SNR (5, 15db). In total, we have 600 (= 3 noise types × 2 SNR × 10 persons × 10 utterances) samples as the test set. For each noise condition, a different set of ten utterances, which are different in text from the test set and taken one by one from all speakers, was used for adaptation of the combination weights and GMM parameters. Namely, the adaptation is performed without depending on speakers and texts.

The frame length is 100ms for amplitude level and ZCR, and 25ms for spectral feature and GMM likelihood. The frame shift is 10ms for all features. Each GMM consists of 32 Gaussians with diagonal covariance matrices, for acoustic parameters of 12 mel-cepstral coefficients with their Δ and Δ-power. The speech GMM is trained with the JNAS (Japanese Newspaper Article Sentences) corpus that includes 32K utterances by 306 speakers. For training of the noise GMM, three types of noise of air conditioner, office and corridor were used. Notice that latter two types of noise are different from those used to make the test set.

3.2. Results

In this paper, we use the method based on GMM only as a reference, because it gives the best performance among the individual methods mentioned in Section 2 in most of the cases[7]. We also investigated the effect of adaptation of GMM parameters. Here, MLLR adaptation is conducted for speech and noise GMMs using the adaptation data of ten utterances, for which correct labels and phonetic transcripts are given. The results are summarized in Tables 1 and 2 for the cases of 5db and 15db, respectively. In these tables, EER in VAD is listed for three noise types. It is observed that the adaptation of GMM is effective consistently. However, the improvement is not so large as the general difference between the proposed method and the GMM-only case. In the proposed method, the gain by the MLLR adaptation is marginal.

Table 3: Number of utterances for weight adaptation (EER(%) in VAD): 5db

	0	1	5	10
AC	11.2	10.8	10.8	10.5
CM	16.8	14.8	14.4	14.4
BS	19.3	18.3	18.1	17.9

Next, we investigate the necessary data size for adapting the combination weights based on MCE training in the proposed method. In Table 3, EER in VAD for the 5db condition is shown by changing the number of utterances up to ten, and "0" means the case where all weights are equal. It is clearly observed that the weight adaptation has a significant impact, and the effect is almost saturated by the first utterance, that is, the adaptation is done by one utterance.

4. Application and Evaluation in ASR

4.1. Utterance Detection

In order to apply the proposed VAD method to ASR, it is necessary to build up an utterance unit based on the frame-wise decision, while rejecting false alarms. Here, we introduce a simple heuristic method: First, we merge segments with a pause gap smaller than a threshold (=100ms). Then, we reject segments shorter than another threshold (=300ms). Here, we assume that spoken utterances must be longer than this threshold. The remaining contiguous segments after these processes are judged as utterances, and fed into an ASR system one by one. Although a more sophisticated method can be explored, it is not the main scope of this work.

4.2. Task and Conditions

For evaluation in ASR, we collected 1345 utterances from the same ten speakers[1], and made a test set by adding the same three types of noise with SNR of 5, 10 and 15db. Thus, we have 12105 samples (= 3 noise types × 3 SNR × 1345 utterances).

The recognition task is simple conversation with a robot. A finite state automaton grammar is handcrafted with a vocabulary of 865 words. The acoustic model is a phonetic tied-mixture (PTM) triphone model based on multi-condition training. These models are fed into our speech recognition engine Julius/Julian.

For the proposed VAD method, weight adaptation was

[1] Precise labels are not necessary in this evaluation.

conducted with the same ten utterances. We chose the threshold value for discriminant function $F(\boldsymbol{X}_t)$ which gave the best EER in the previous Section for each condition.

4.3. Results

In Tables 4∼6, ASR performance in word accuracy is listed for the SNR conditions of 5db, 10db and 15db, respectively. In these tables, the proposed method is compared against four individual methods. The ZCR-based method did not work at all for the craft-machine noise (CM) which has frequent zero crossing. It is clearly seen that the proposed method outperforms the other methods in almost all conditions, and realizes significant improvement on average.

For reference, the tables give the accuracy obtained when the weight adaptation is conducted with the test set "(Closed)", and also when the utterance segmentation (VAD) is done manually "(Oracle)". The proposed method shows comparable performance to the "Closed" case. It means that the weight adaptation is reliably performed without depending on the given data. When compared with the "Oracle", larger degradation is observed for the cases of lower SNR and background speech (BS). The result confirms the significance of VAD and difficulty in non-stationary noise such as background speech. In actual, we observed many false alarms causing insertion errors in ASR in this noise.

Table 4: Word accuracy in ASR (%): 5db

	AC	CM	BS	average
Amplitude	44.4	16.8	-6.6	18.2
ZCR	25.6	NA	-54.7	-14.6
Spectrum	34.9	14.2	-4.3	15.0
GMM	54.7	-14.2	-6.4	11.4
Proposed	57.2	32.9	3.8	31.3
(cf) Closed	54.2	37.7	0.7	30.9
(cf) Oracle	70.8	55.5	62.3	62.9

Table 5: Word accuracy in ASR (%): 10db

	AC	CM	BS	average
Amplitude	62.9	45.8	29.4	46.1
ZCR	58.2	NA	-13.7	22.2
Spectrum	54.2	49.5	25.5	43.1
GMM	72.3	37.6	29.3	46.4
Proposed	75.1	59.7	37.8	57.5
(cf) Closed	74.5	63.1	37.6	58.4
(cf) Oracle	80.9	70.4	79.0	76.8

Table 6: Word accuracy in ASR (%): 15db

	AC	CM	BS	average
Amplitude	71.2	63.3	48.5	61.0
ZCR	78.6	NA	25.7	52.1
Spectrum	65.3	68.4	46.8	60.2
GMM	76.2	62.9	46.1	61.7
Proposed	77.8	74.0	54.4	68.7
(cf) Closed	77.5	74.5	55.5	69.2
(cf) Oracle	82.2	79.2	83.3	81.5

AC: air conditioner, CM: craft machine, BS: background speech

5. Conclusion

We have presented a robust VAD method by adaptively combining the four different features. In the experimental evaluations with a variety of noise conditions, the proposed method realizes the significantly better performance than the conventional individual techniques. It is also shown that the weight adaptation is possible with only one utterance and as reliable as in the closed training. In the future, we will investigate on-line adaptation of the weights and enhancement for non-stationary noise.

6. References

[1] K.Li, M.Swamy, M.Omair, "An improved voice activity detection using higher order statistics," *IEEE Trans. ASSP.*, vol.13, no.5, pp.965-974, 2005.

[2] ITU, "A silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70," *ITU-T Recommendation G.729-Annex B*, 1996.

[3] 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,"*Proc. ICASSP*, vol.II, pp.1093-1096, 2004.

[4] P.N.Garner, T.Fukada, Y.Komori, "A Differential spectral voice activity detector," *Proc. ICASSP*, vol.I, pp.597-600, 2004.

[5] A.Lee, K.Nakamura, R.Nishimura, H.Saruwatari, K.Shikano, "Noise robust real world spoken dialog system using GMM based rejection of unintended inputs," *Proc. ICSLP*, vol.I, pp.173-176, 2004.

[6] J.Sohen, N.Kim, W.Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol.16, no.1, pp.1-3, 1999.

[7] Y.Kida, T.Kawahara, "Voice activity detection based on optimally weighted combination of multiple features," *Proc. INTERSPEECH*, pp.2621–2624, 2005.

[8] T.Yamada, J.Okada, K.Takeda, N.Kitaoka, M.Fujimoto, S.Kuroiwa, K.Yamamoto, T.Nishiura, M.Mizumachi, S.Nakamura, "Integration of noise reduction algorithms for Aurora2 task," *Proc. EUROSPEECH*, pp.1769-1772, 2003.