DUAL-MICROPHONE VOICE ACTIVITY DETECTION BASED ON USING OPTIMALLY WEIGHTED MAXIMUM *A POSTERIORI* PROBABILITIES

Seng Hyun Huang, Jihwan Park, and Joon-Hyuk Chang

Hanyang University
Department of Electronics Engineering
Seoul, Korea

ABSTRACT

In this paper, we propose to improve the dual-microphone voice activity detection (VAD) technique for which a discriminative weight training is applied to achieve optimally weighted spatial features. In our approach, we first derive the maximum *a posteriori* (MAP) probabilities from the spatial features such as the power level difference ratio (PLDR), phase vector, and coherence. Then, we combine each MAP probability within the minimum classification error (MCE) framework to offer an optimal VAD decision in a spectral domain. Experimental results show that the proposed dual-microphone VAD algorithm shows better performances than the conventional dual-microphone VAD methods, which solely utilize the PLDR, phase, and spectral coherence.

Index Terms— Voice activity detection, dual-microphone, discriminative weight training, minimum classification error

1. INTRODUCTION

Voice activity detection (VAD) in a speech signal plays a crucial role in speech coding, speech recognition, and speech enhancement. Traditionally, the statistical-model based VAD employing the decision-directed (DD) method parameter estimation and reported high detection accuracy. The superiority of the statistical model-based VAD has been recognized in most studies in which the likelihood ratio (LR) test is derived given a set of hypotheses, [1]. The statistical model-based VAD was further improved by adopting the minimum classification error (MCE) algorithm [2] in which optimally weighted LRs are integrated into the VAD decision.

One of the predominant dual-microphone VAD techniques employs the coherence technique [3], [4], which is based on the assumption that the speech signals in two channels are correlated, which the noisy signals are relatively uncorrelated. The technique studied by Arabi and *et al*. [5] devised the VAD method to employ the time difference of arrival (TDOA) of input signals at the two microphones. In addition, the power level difference (PLD) was developed in [7] at which the basic concept of the PLD relies on the fact that speech signals have different power levels between microphones, while the power levels of noise signals are almost equivalent. Then, Choi and Chang [8] presented a novel algorithm to incorporate the PLD of noise during speech pauses and then proposed the two-step PLD ratio (denoted by PLDR), which is the ratio of the PLD of speech and noise estimated during noise periods. Specifically, the long-term power level difference ratio (LT-PLDR) and short-term power level difference ratio (ST-PLDR) were computed to characterize the long-term evolution and short-term variation, respectively.

In this paper, we propose a novel dual-microphone VAD technique using optimally weighted spatial features. In addition to the PLDR proposed in [8], we consider the more spatial features such as the phase vector and coherence and apply the MCE scheme in an attempt to represent the different contributions of the spatial features for the VAD. Above all, the maximum *a posteriori* (MAP) probabilities are first derived from each feature based on the model-trust minimizing algorithm to classify the speech presence or absence regions. Then, the optimal weights are achieved by the generalized probabilistic descent (GPD) technique and applied to the each MAP probability to be optimally adjusted in the VAD decision rule. The performance of the proposed algorithm is evaluated by extensive objective tests under various acoustic conditions. Based on a number of experiments, the proposed VAD technique combining several models is superior and found to yield a better performance than solely utilize the feature under various acoustical circumstances.

2. PROPOSED DUAL-MICROPHONE VOICE ACTIVITY DETECTION USING THE MCE ALGORITHM

It is assumed that input signals at the two microphones are denoted $y_i(t) = x_i(t) + n_i(t)$, which is the sum of a clean speech signal $x_i(t)$ and a noise signal $n_i(t)$. By taking the

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 20141A2A1A10049735). This research was also supported by LG Electronics.

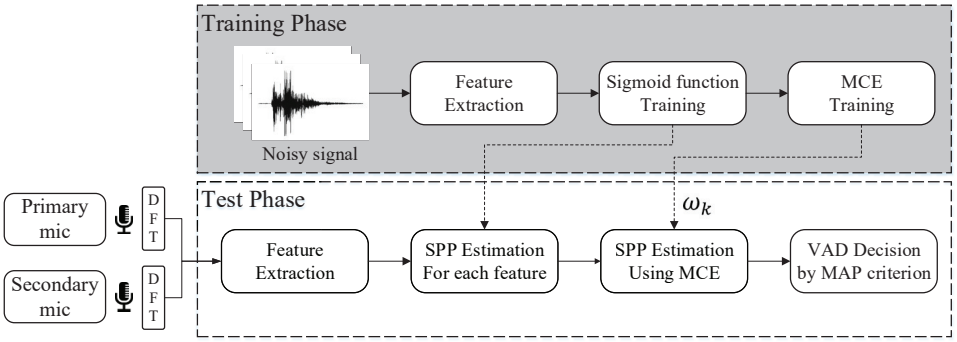

Fig. 1. Overall block diagram of the proposed two-microphone VAD approach

discrete Fourier transform (DFT) of the noisy signal $y_i(t)$, then the equation in the time-frequency domain is obtained by

$$Y_i(k,n) = X_i(k,n) + N_i(k,n). \qquad i = 1,2 \tag{1}$$

Given two hypotheses, $H_0(k,l)$ and $H_1(k,l)$, which indicate speech absence and presence, respectively. Then, it is assumed that

$$\begin{aligned} H_0(k,l) &: Y_i(k,n) = N_i(k,n) \\ H_1(k,l) &: Y_i(k,n) = X_i(k,n) + N_i(k,n). \end{aligned} \tag{2}$$

At first, the PLDR VAD technique was found to improve the performance of the VAD which is ratio of the PLD $\Delta P_Y(k,n)$ and the PLD of the noise $\widehat{\Delta P}_N(k,n)$ in [2] as follows:

$$Q(k,n) = \frac{\widehat{\Delta P}_Y(k,n)}{\widehat{\Delta P}_N(k,n)} \tag{3}$$

where the PLD of the noise $\widehat{\Delta P}_N(k,n)$ is obtained by using the minima controlled recursive averaging (MCRA) [9] scheme. Note that the PLDR is composed of the LT-PLDR and ST-PLDR, which are used to characterize the long-term evolution and the short-term variation, respectively. However, we believe that it is not enough for the dual-microphone VAD since it does not consider additional spatial benefits such as spatial correlation as well as the phase difference. For this reason, we develop the combined VAD decision by using multiple spatial features such as the PLDR, phase vector, and spectral coherence at which the optimal weights are assigned to the multiple features by a help of the MCE method.

Specifically, based (2), the coherence function [3], [4] which represent by a correlation of a signal is given as follows:

$$\Gamma_{Y_1Y_2}(k,n) = \frac{P_{Y_1Y_2}(k,n)}{\sqrt{P_{Y_1}(k,n)P_{Y_2}(k,n)}} \tag{4}$$

where P_{Y_1} and P_{Y_2} are the PSD of two microphones, $P_{Y_1Y_2}$ denotes the cross power spectral density (CPSD), respectively. On the other hand, the phase vector [6] which use the phase difference between the two-microphone is given by

$$a(k,n) \triangleq \left[\frac{\acute{q}_1(k,n)}{|\acute{q}_1(k,n)|}\right]^{\mathrm{T}} \tag{5}$$

where $\acute{q}_1(k,n)$ is normalized by the first element of the principal eigenvector [10] which has the largest eigenvalue. The *a posteriori* probability of each feature is obtained by using the sigmoid fitting approach [11] as follows:

$$p(H(n) = H_1|\phi_i(n)) = \frac{1}{1+\exp(a\phi_i(n)+b)} \tag{6}$$

where ϕ_i is the each feature and a and b are the slope parameter and bias parameter of each feature, respectively. Based on this, the dual-microphone VAD is proposed using multiple spatial features by defining the optimally weighted *a posteriori* probability as given by

$$\Lambda_\omega(n) = \sum_{i=1}^{N} \omega_i p(H(n) = H_1|\phi_i(n)) \tag{7}$$

where $\{\omega_i\}$ are weights for the MAP probabilities i.e., $p(H(n) = H_1|\phi_i(n))$ and N denotes the total number of features. The weights $\{\omega_i\}$ should satisfy the following conditions.

$$\sum_{i=1}^{M} \omega_i = 1, \;\; \omega_i > 0 \text{ for } i = 1,2,\cdots,N. \tag{8}$$

Note that $\mathbf{\Lambda}_\omega(n)$ represents the optimally weighted feature vector in our approach. In time, two discriminant functions of speech (g_s) and noise (g_n) classify to decide if each frame is classified into speech or noise according to the following conditions.

$$g_s(\mathbf{\Lambda}_\omega(n)) = \Lambda_\omega(n) - \theta \tag{9}$$

$$g_n(\mathbf{\Lambda}_\omega(n)) = \theta - \Lambda_\omega(n) \tag{10}$$

where θ is the threshold value of the combined score. From the combined score, we estimate the weight for which the features are differently contributed in classifying speech or noise. Subsequently, the weights are found by the discriminative weight training as follows:

$$\mathcal{D}(\mathbf{\Lambda}_\omega(n)) = \begin{cases} -g_s(\mathbf{\Lambda}_\omega(n)) + g_n(\mathbf{\Lambda}_\omega(n)), \text{if } g_s \text{ is true} \\ -g_n(\mathbf{\Lambda}_\omega(n)) + g_s(\mathbf{\Lambda}_\omega(n)), \text{if } g_n \text{ is true} \end{cases} \tag{11}$$

where $\mathcal{D}(\mathbf{\Lambda}_\omega(n))$ is the misclassification measure of training data $\{\mathbf{\Lambda}_\omega(n)\}$. If the classification is correct, $\mathcal{D}(\mathbf{\Lambda}_\omega(n))$ is negative, which raises the error. Specifically, the GPD technique approximates the empirical classification error by a smooth objective function which is the zero to one step loss function with a gradient γ of the sigmoid function as given by

$$L(t) = \frac{1}{1 + \exp(-\gamma \mathcal{D}(\mathbf{\Lambda}_\omega(n)))}, \quad \gamma > 0 \tag{12}$$

where the loss function yields a minimum value when the weights are optimized. To consider the condition in (8), the following parameter transformation is applied.

$$\tilde{\omega}_i = \log \omega_i. \tag{13}$$

Then, the weights of each frequency bin, $\tilde{\omega}_k$ is updated based on the steepest descent algorithms as given by

$$\tilde{\omega}_i(n+1) = \tilde{\omega}_i(n) - \epsilon \frac{\partial L(t)}{\partial \tilde{\omega}_i}|_{\tilde{\omega}_i = \tilde{\omega}_i(n)} \tag{14}$$

where ϵ is a step size. Once $\tilde{\omega}_i$ is updated, we adopt the inverse transform to $\tilde{\omega}_i$ as follows:

$$\omega_i = \frac{\exp(\tilde{\omega}_i)}{\sum_{j=1}^{M} \exp(\tilde{\omega}_i)}. \tag{15}$$

Finally, we perform the VAD decision based on the MAP technique by using the MCE training as follows:

$$\frac{p(H(n) = H_1 | \Phi(n))}{p(H(n) = H_0 | \Phi(n))} \gtrless_{H_0}^{H_1} \eta \tag{16}$$

where η is a given threshold.

3. EXPERIMENTS

3.1. Experiment setup

We evaluated our approach to estimate the proposed dual-microphone VAD using the optimally weighted features with objective measures under various conditions. The performance of the proposed algorithm was compared with traditional VAD techniques consisting of the single feature such as the LT-PLDR ($\mathcal{L}$), ST-PLDR ($\mathcal{S}$), phase vector ($\mathcal{P}$), and coherence function ($\mathcal{C}$). The evaluation was conducted by different four types of combination using the MCE scheme

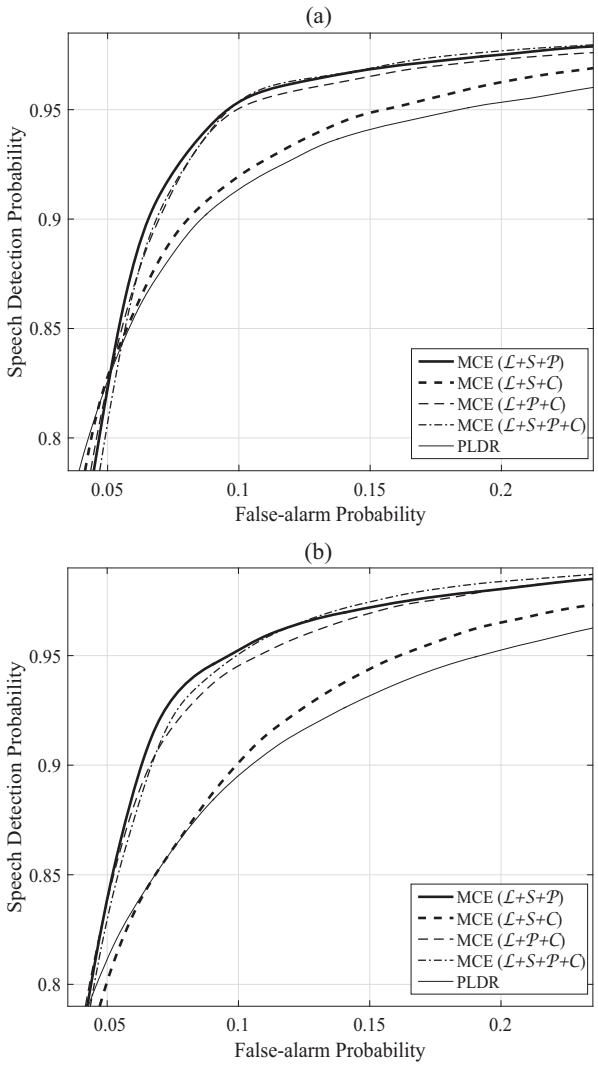

Fig. 2. ROC curves for various noise environments with approx. 6 dB SNR. (a) babble noise (b) office noise

that MCE ($\mathcal{L}$ + $\mathcal{S}$ + $\mathcal{P}$), MCE ($\mathcal{L}$ + $\mathcal{S}$ + $\mathcal{C}$), MCE ($\mathcal{L}$ + $\mathcal{P}$ + $\mathcal{C}$), and MCE ($\mathcal{L}$ + $\mathcal{S}$ + $\mathcal{P}$ + $\mathcal{C}$)

For objective evaluation, the efficient metrics that speech hit rate (P_{sh}) and non-speech hit rate (P_{nh}) were employed [8]. For the training and test phase, noisy sentences were recorded at various distances of 1 m, 3 m, and 5 m and at azimuth angles of 0°, 90°, and 180° between the speech source at the dummy head and the noise source. For simulating noisy environments, speech data was artificially contaminated with four different noisy sources such as babble, office, white, and factory from the NOISEX-92 database [13]. The total samples were composed of 520 s long speech data which was manually labeled the speech and non-speech segments of the speech signal every 10 ms frame. The proportion of the hand-marked active speech frame was 57.1 %, which consist of 44.5 % voiced sounds and 13.4 % unvoiced sounds.

Table 1. Comparison of the conventional VAD methods and the proposed techniques with approx. 6 dB SNR

Source Location	Noise Environments	Babble		Office		White		Factory	
		P_{sh}	P_{nh}	P_{sh}	P_{nh}	P_{sh}	P_{nh}	P_{sh}	P_{nh}
0°	PLDR [8]	93.44	89.95	93.41	89.23	95.29	90.44	91.81	89.41
	Phase vector [6]	92.95	87.47	94.52	87.88	62.9	74.19	88.25	87.57
	Coherence [3]	91.95	85.86	91.60	84.52	82.01	93.32	87.86	84.73
	MCE ($\mathcal{L}+\mathcal{S}+\mathcal{P}$)	**95.97**	**89.39**	**96.25**	**90.04**	94.71	90.12	**94.10**	**89.45**
	MCE ($\mathcal{L}+\mathcal{S}+\mathcal{C}$)	94.56	89.50	95.38	87.63	**96.03**	**90.05**	94.56	86.99
	MCE ($\mathcal{L}+\mathcal{P}+\mathcal{C}$)	96.88	87.68	95.88	89.62	93.40	90.07	94.31	88.37
	MCE ($\mathcal{L}+\mathcal{S}+\mathcal{P}+\mathcal{C}$)	96.69	87.70	96.09	89.50	92.92	89.83	98.15	82.98
90°	PLDR [8]	93.83	89.62	93.21	89.54	94.70	89.84	91.90	89.69
	Phase vector [6]	95.60	88.66	94.96	89.91	84.47	86.76	85.30	88.68
	Coherence [3]	90.85	85.65	89.82	85.31	80.75	91.75	87.71	81.83
	MCE ($\mathcal{L}+\mathcal{S}+\mathcal{P}$)	**96.17**	**88.89**	**95.99**	**89.50**	93.42	89.96	**94.39**	**89.29**
	MCE ($\mathcal{L}+\mathcal{S}+\mathcal{C}$)	95.66	88.76	94.93	88.44	**95.43**	**89.75**	94.03	88.39
	MCE ($\mathcal{L}+\mathcal{P}+\mathcal{C}$)	96.45	88.42	96.14	89.35	92.09	90.55	94.59	88.75
	MCE ($\mathcal{L}+\mathcal{S}+\mathcal{P}+\mathcal{C}$)	96.33	88.38	96.33	89.09	91.86	89.97	94.59	88.38
180°	PLDR [8]	89.04	87.04	89.44	86.39	78.17	90.53	86.48	85.60
	Phase vector [6]	74.83	72.21	88.41	85.15	79.61	49.11	70.62	80.79
	Coherence [3]	80.93	83.64	78.35	84.61	76.02	63.91	73.91	79.12
	MCE ($\mathcal{L}+\mathcal{S}+\mathcal{P}$)	**90.00**	**87.23**	93.59	86.89	**83.29**	**85.78**	**89.42**	**86.76**
	MCE ($\mathcal{L}+\mathcal{S}+\mathcal{C}$)	90.74	85.98	90.41	85.83	80.48	89.34	86.83	86.24
	MCE ($\mathcal{L}+\mathcal{P}+\mathcal{C}$)	88.49	87.57	93.12	87.47	80.83	88.07	87.10	88.97
	MCE ($\mathcal{L}+\mathcal{S}+\mathcal{P}+\mathcal{C}$)	88.07	87.24	**93.04**	**87.57**	82.20	85.06	87.01	88.72

3.2. Experimental Results

We evaluated the performance of the proposed approach compared with the traditional dual-microphone VAD techniques [3], [6], and [8]. Then, for evaluating the detection accuracy in terms of the speech and non-speech segment, noise source was located at 0°, 90°, and 180°. Then, the proposed method was evaluated under the various conditions that four noise types at different distances like an 1 m, 3 m, and 5 m. The results of this experiment is summarized in Table 1. According to this table, it was found that the proposed dual-microphone VAD technique using multiple features was superior to the conventional dual-microphone VAD techniques for all tested conditions. In particular, the MCE ($\mathcal{L}+\mathcal{S}+\mathcal{P}$) showed the best performance in terms of the probability of the detection for speech, especially babble, office, and factory noises. Note that the phase vector in the MCE ($\mathcal{L}+\mathcal{S}+\mathcal{P}$) is relatively attractive especially when the distance was short. As this tendency was noticeably observed at the 90° and 180° azimuth. In this result, the MCE ($\mathcal{L}+\mathcal{S}+\mathcal{P}$) was outstanding against traditional VAD techniques in all tested conditions. Especially, the results for office and destroyer engine noise environments outperformed all other algorithms in four noise conditions. Also, the receiver operating characteristics (ROCs), showing the trade-off between speech detection probability and false-alarm probability of babble, office, white, and factory noise environments are shown in Fig. 2. As a result, the proposed vad technique using multiple features showed the performance improvement compared to VAD techniques using the solely feature from dual-microphone. It is obvious that the result for the detection probability using the optimally weighted features to multiple features is considerably improved for babble and office noise environments.

4. CONCLUSION

In this paper, we proposed a dual-microphone VAD technique using optimally weighted spatial features including the PLDR, coherence, and phase vector. The proposed VAD technique using the MCE framework adopt the optimal weights for spatial features to the VAD algorithm by discriminative weight training. First, the MAP probability of the traditional VADs is estimated by model-trust algorithm. Then, the MCE training is adopted to obtain the optimal weights for each spatial features. We apply the MCE scheme for all of the combination spatial features to evaluate the performance of VAD techniques. Our experimental results showed that the proposed VAD technique using multiple spatial features provides reliable VAD performances under various noise environments including non-stationary noise conditions that babble and office noises.

5. REFERENCES

[1] J. Sohn, N. S. Kim, and W. sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1-3, Jan. 1999.

[2] S.-I. Kang, Q.-H. Jo, and J.-H. Chang, "Discriminative weight training for a statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 15, pp. 170-173, Jan. 2008.

[3] R. Le Bouquin-Jeanns and G. Faucon, "Study of a voice activity detector and its influence on a noise reduction system," *Speech Commun.*, vol. 16, pp. 245-254, Apr. 1995.

[4] N. Yousefian and P. C. Loizou, "A dual-microphone speech enhancement algorithm based on the coherence function," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 2, pp. 599-609, Feb. 2012.

[5] P. Aarabi and G. Shi, "Phase-based dual-microphone robust speech enhancement," *IEEE Trans. Systems, Man, Cybern. B*, vol. 34, no. 4, pp. 1763-1773, Aug. 2004.

[6] G. Kim and N. I. Cho, "Voice activity detection using phase vector in microphone array," *Electronics Lett.*, vol. 43, no. 14, pp. 783-784, Jul. 2007.

[7] N. Yousefian, A. Akbari, and M. Rahmani, "Using power level difference for near field dual-microphone speech enhancement," *Appl. Acoust.*, vol. 70, no. 11-12, pp. 1412-1421, Dec. 2009.

[8] J.-H. Choi and J.-H. Chang, "Dual-microphone voice activity detection technique based on two-step power level difference ratio," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 6, Jun. 2014.

[9] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Process. Lett.*, vol. 9, no. 1, pp. 12-15, Jan. 2002.

[10] Yang. B, "Projection approximation subspace tracking," *IEEE Signal Process. Lett.*, vol. 43, no. 1, pp. 95-107, Jan. 1995.

[11] J.-H. Chang, Q.-H. Jo, D. K. Kim, and N. S. Kim, "Global soft decision employing support vector machine for speech enhancement," *IEEE Signal Process. Lett.*, vol. 16, no. 1, pp. 57-60, Jan. 2009.

[12] Y. Kida and T. Kawahara, "Voice activity detection based on optimally weighted combination of multiple features," in *Proc. Interspeech*, pp. 2621-2624, Sep. 2005,

[13] A. Varga and H. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247-251, Jul. 1993.