Dual-Microphone Voice Activity Detection Based On Using Optimally Weighted Maximum *A Posteriori* Probability

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

ABSTRACT

In this study, 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. The maximum a posteriori (MAP) probabilities from the spatial features are combined using the minimum classification error (MCE) framework to offer an optimal VAD decision in a spectral domain.

Keywords – voice activity detection, dual-microphone, discriminative weight training, minimum classification error

INTRODUCTION

- Motivation behind our approach is to use the most spatial information available from the two microphones, which successfully characterizes the dynamic evolution of speech in time especially in the non-stationary noise environments.
- We consider not only single spatial features, but multiple spatial features such as power level difference ratio (PLDR), coherence function, and phase difference by applying MCE scheme.
- We attempt to incorporate the different contributions of the spatial features under dynamic acoustic environments by applying the MCE scheme.

PROPOSED METHOD

- **Feature selection**

- PLDR : PLDR is the ratio of the power level difference (PLD) and the PLD of the noise which is consist of the long-term PLDR (LT-PLDR : $\mathcal{L}$) and the short-term PLDR (ST-PLDR : $\mathcal{S}$).

$$Q(k,n) = \frac{\widehat{\Delta P}_Y(k,n)}{\widehat{\Delta P}_N(k,n)} \quad \text{where}$$

$\widehat{\Delta P}_y$: PLD of the signal
$\widehat{\Delta P}_n$: PLD of the noise
k : frequency bin
n : frame index

- Coherence function ($\mathcal{C}$): coherence function represent by a correlation of a signal

$$\Gamma_{Y_1Y_2}(k,n) = \frac{P_{Y_1Y_2}(k,n)}{\sqrt{P_{Y_1}(k,n)P_{Y_2}(k,n)}} \quad \text{where}$$

$P_{Y_1Y_2}$: cross power spectral density (CPSD)
$P_{Y_1}P_{Y_2}$: PLD of two microphones
k : frequency bin
n : frame index

- Phase vector ($\mathcal{P}$): Phase difference represent the phase difference of the signal.

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

$\acute{q}_1$: first element of the principal eigenvector
k : frequency bin
n : frame index

- ***A posteriori* probability of each feature**

The a posteriori probability of each feature is obtained by using the sigmoid fitting approach which of training by the model-trust algorithm to minimize cross-entropy error function as follows:

$$p(H(n) = H_1|\phi_i(n)) = \frac{1}{1+\exp(a\phi_i(n)+b)} \quad \text{where}$$

ϕ : spatial feature value
i : feature index
a : slope parameter
b : bias parameter

- **Dual-Microphone VAD using multiple spatial features**

The dual-microphone VAD is proposed using multiple spatial features by defining the optimally weighted a posteriori probability as given by

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

$\{\omega_i\}$: weights for the MAP probabilities
N : total number of features

Note that Λ_ω(n) represents the optimally weighted feature vector in our approach. Then, two discriminant functions of speech and noise classify to decide each frame state from combined score as given by

$$g_s(\mathbf{\Lambda}_\omega(n)) = \Lambda_\omega(n) - \theta$$
$$g_n(\mathbf{\Lambda}_\omega(n)) = \theta - \Lambda_\omega(n) \quad \text{where} \quad \theta : \text{threshold value}$$

From the combined score, we estimate the weight for which the features are differently contributed in classifying speech. Subsequently, the weights are found by the discriminative weight training as follows:

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

where $\mathcal{D}(\mathbf{\Lambda}_\omega(n))$: misclassification measure of training data

Specifically, the GPD technique approximates the empirical classification error by a smooth objective function which is the step loss function of the sigmoid function as given by

$$L(t) = \frac{1}{1+\exp(-\gamma\mathcal{D}(\mathbf{\Lambda}_\omega(n)))}, \quad \gamma > 0 \qquad \text{where} \quad \gamma : \text{gradient}$$

where the loss function yields a minimum value when the weights are optimized. Then, the weights of each features are updated as follows:

$$\tilde{\omega}_i = \log \omega_i$$

$$\tilde{\omega}_i(n+1) = \tilde{\omega}_i(n) - \epsilon \frac{\partial L(t)}{\partial \tilde{\omega}_i}\Big|_{\tilde{\omega}_i = \tilde{\omega}_i(n)} \qquad \text{where} \quad \epsilon : \text{step size}$$

Once $\tilde{\omega}_i$ is updated, we adopt the inverse form to $\tilde{\omega}_i$ as given by

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

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

$$\frac{p(H(n) = H_1|\Phi(n))}{p(H(n) = H_0|\Phi(n))} \gtrless_{H_0}^{H_1} \eta \qquad \text{where} \quad \eta : \text{threshold}$$

- **Overall block diagram**

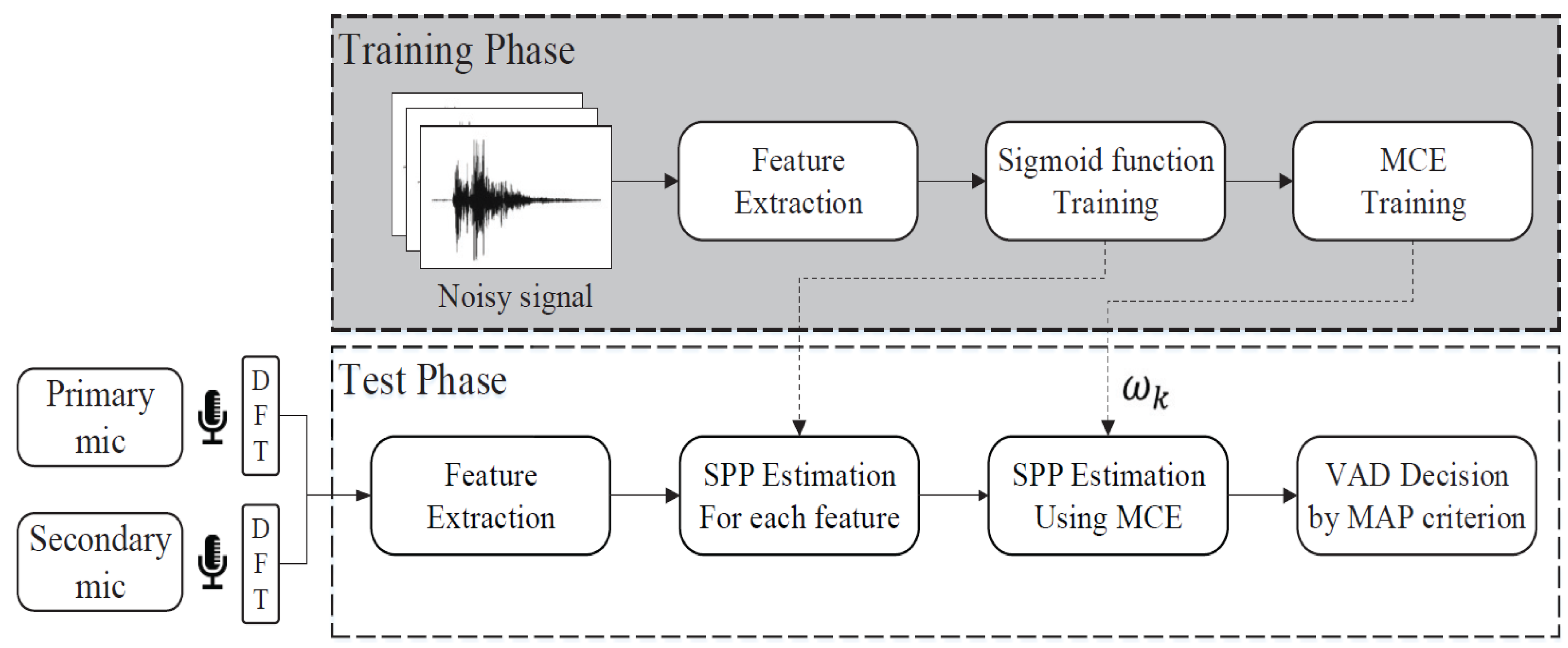

Fig. 1. Overall block diagram of the proposed two-microphone VAD approach

RESULTS

- **Experimental environment**

– The total samples were composed of 520s long speech data and noisy sentences were recorded at various distances and azimuth angles.

– 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.

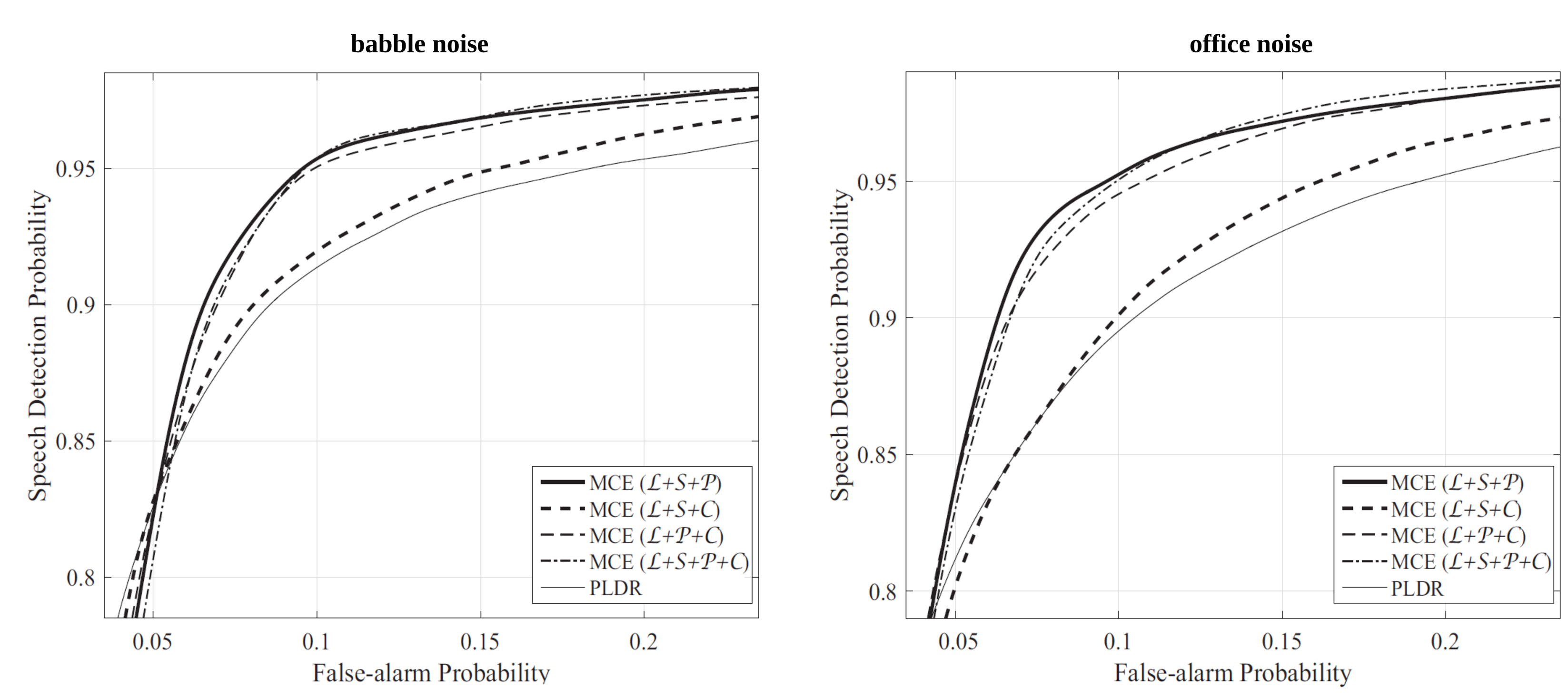

Fig. 2. ROC curves for various noise environments with approx. 6 dB SNR

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

- PLDR : 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.
- Phase vector : 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.
- Coherence function : 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.

CONCLUSIONS

- In this study, we proposed a dual-microphone VAD technique using optimally weighted spatial features including the PLDR, coherence, and phase vector.
- The principal contribution is using the MCE framework adopt the optimal weights for spatial features to the VAD algorithm by discriminative weight training.
- To optimize the weights of multiple spatial features, 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.
- The proposed VAD technique using multiple spatial features provides reliable VAD performances under various noise environments including non-stationary conditions that babble and office noises.