INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Improved *a priori* SAP estimator in complex noisy environment for dual channel microphone system

Youna Ji, Young-cheol Park

Computer and Telecomm. Eng. Division
Yonsei University Wonju, Korea
jyn282@yonsei.ac.kr

Abstract

In this paper, *a priori* speech absence probability (SAP) estimator is proposed for accurately obtaining the speech presence probability (SPP) in a complex noise field. Unlike previous techniques, the proposed estimator considers a complex noise sound field where the target speech is corrupted by a coherent interference with diffuse noise around. The proposed algorithm estimates *a priori* SAP based on the normalized speech to interference plus diffuse noise ratio (SINR) being expressed in terms of the speech to interference ratio (SIR) and the directional to diffuse noise ratio (DDR). The SIR is obtained from a quadratic equation of the magnitude-squared coherence (MSC) between two microphone signals. A performance comparison with several advanced *a priori* SAP estimators was conducted in terms of the receiver operating characteristic (ROC) curve. The proposed algorithm attains a correct detection rate at a given false-alarm rate that is higher than those attained by conventional algorithms.

Index Terms: speech presence probability, *a priori* speech absence probability, complex noisy environment

1. Introduction

The speech presence probability (SPP) is crucial for many applications, such as speech enhancement and dereverberation. A general SPP estimator can be derived under the assumption that the spectral coefficients of speech and noise can be modelled as complex Gaussian random variables [1], and is often computed using *a priori* and *a posteriori* SNRs, as well as *a priori* speech absence probability (SAP) [2]. Theoretically, *a priori* SAP does not depend on observation, and so, it can be set as a fixed value [1, 3, 2]. However, in practice, the SAP varies with time and frequency, depending on the words spoken [2]. Hence, it is more appropriate to estimate the *a priori* SAP in each time-frequency (TF) unit instead of using a fixed value.

Several algorithms have been proposed to estimate and update *a priori* SAP. In [4, 5], a soft decision approach was proposed using an assumption that the neighboring frequency bins of consecutive frames in the speech presence region were highly correlated. In this approach, the estimated SNRs were averaged and mapped onto a value between zero to one for use as an estimate of SAP. Although SNR is strongly correlated with SAP, the accuracy of the obtained SAP is highly affected by the noise estimation performance and the mapping function. Meanwhile, a multichannel *a priori* SAP estimator was proposed in [6], where an estimate of the directional to diffuse ratio (DDR) was utilized. Since this estimator does not require the statistical information of the noise or speech, the detection accuracy is decoupled from the noise estimation performance. A more recent study [7] proposed an algorithm for determining *a priori* SAP based on magnitude-squared coherence (MSC) for a dual-channel microphone system. This algorithm utilizes the real and imaginary parts of the coherence function between the input signals as a criterion for estimating the normalized SNR [8, 9]. Thus, the algorithm requires neither an additional mapping function nor prior knowledge of the noise or speech statistics.

It should be noted that most conventional SPP estimators had been used assuming a diffuse noise field with one single directional target signal. In contrast, little attention has been paid to complex acoustic noise fields existing together with background diffuse noise that propagates from all directions with equal amplitude and a random phase [10]. The diffuse noise exhibits a high correlation at low frequencies and a very low coherence over the remaining frequency spectrum [10, 2]. Interference is a noise source incident from a specific direction and is often highly correlated over most of the frequency components. The above conventional SPP estimation techniques do not consider the coherent interference, and so, produce biased results.

In this paper, *a priori* SAP estimator in [7] is extended to a complex acoustic environment in an effort to improve the performance of the SPP estimator in a more practical and complex noise field. In the proposed algorithm, the *a priori* SAP is computed using two parameters: DDR and the speech to interference ratio (SIR). It is first shown that the normalized DDR and SIR can be numerically expressed using the MSC of two-channel observation signals. Computer simulations were conducted in a situation where a frontal target speech was corrupted by a directional interference with a surrounding diffuse noise.

2. General SPP estimation in dual channel system

The observation signals of the dual-channel microphone system can be represented in the frequency domain as

$$Y_i(k,l) = S_i(k,l) + N_i(k,l), i = 1,2, \qquad (1)$$

where $S_i(k,l)$ is the target speech at the ith-channel microphone. k and l denote the frequency bin and frame indices, respectively. $N_i(k,l)$ are observed noise signals, which can be either white or colored, but are uncorrelated with $S_i(k,l)$. The observed signals can be written in vector notation as $\mathbf{y}(k,l) = [Y_1(k,l), Y_2(k,l)]^T$, and the power spectral density(PSD) matrix of $\mathbf{y}(k,l)$ is defined as $\mathbf{\Phi}_{yy}(k,l) = E[\mathbf{y}(k,l)\mathbf{y}^H(k,l)]$.

Let us assume that $H_1(k,l)$ and $H_0(k,l)$ are two-state hypotheses, which represent speech presence and absence, respectively. Then, under the assumption that the desired speech and

http://dx.doi.org/10.21437/Interspeech.2016-894

noise components are modeled as complex multivariate Gaussian random variables, we obtain the multichannel *a posteriori* SPP estimate as in [1]:

$$p\,(k,l) = P[H_1(k,l)|\mathbf{y}(k,l)] \\ = \left\{1+\frac{q(k,l)}{1-q(k,l)}[1+\xi(k,l)]\text{exp}\left[-\frac{\beta(k,l)}{1+\xi(k,l)}\right]\right\}^{-1}, \quad (2)$$

where $\xi(k,l) = tr[\mathbf{\Phi}_{nn}^{-1}(k,l)\mathbf{\Phi}_{xx}(k,l)]$ denotes the *a priori* SNR, $q(k,l) = P[H_0(k,l)]$ is the *a priori* SAP, and $\beta(k,l) = \mathbf{y}^H(k,l)\mathbf{\Phi}_{nn}^{-1}(k,l)\mathbf{\Phi}_{xx}(k,l)\mathbf{\Phi}_{nn}^{-1}(k,l)\mathbf{y}(k,l)$.

3. Proposed *a priori* SAP estimator in complex noisy environment

3.1. Signal model in complex noisy environment

In an environment with complex noise, the dual channel input signals can be represented in the frequency domain as

$$Y_i(k,l) = S_i(k,l) + V_i(k,l) + \tilde{N}_i(k,l), i = 1,2, \quad (3)$$

where $V_i(k,l)$ is a directional interference and $\tilde{N}_i(k,l)$ is the diffuse noise signal. So, the noise in (1) is sum of the interference and the diffuse noise, i.e., $N_i(k,l) = V_i(k,l) + \tilde{N}_i(k,l)$. The interference is a dominant directional noise, that is, the dominant noise incident from a specific direction. Thus, it is highly correlated with the two microphones signals. On the other hand, diffuse noise, which is propagating in all directions simultaneously with equal power and random phase, is often uncorrelated with each other except at low frequency [10, 11].

The coherence between the two microphone signals can be calculated as

$$\Gamma_Y(k,l) = \frac{\Phi_{YY}^{12}(k,l)}{\sqrt{\Phi_{YY}^{11}(k,l)\Phi_{YY}^{22}(k,l)}}, \quad (4)$$

where $\Phi_{YY}^{ij}(k,l) = E\{Y_i(k,l)Y_j^*(k,l)\}, i,j = 1,2$ are cross- and auto-PSDs of the microphone signals. For the algorithm development, we assume that, $E\{S_i(k,l)V_i^*(k,l)\} = 0$, and $E\{S_i(k,l)\tilde{N}_i^*(k,l)\} = 0$. We will omit the frequency and frame indices whenever necessary.

According to the study in [7], the noisy coherence can be the represented weighted sum of the directional signal and the diffuse noise coherences:

$$\Gamma_Y(k,l) = \Gamma_D(k,l)\left(\sqrt{\frac{DDR_1}{1+DDR_1}\cdot\frac{DDR_2}{DDR_2+1}}\right) \\ +\Gamma_N(k,l)\left(\sqrt{\frac{1}{DDR_1+1}\cdot\frac{1}{DDR_2+1}}\right), \quad (5)$$

where $\Gamma_D(k,l)$ and $\Gamma_N(k,l)$ are the coherences of the directional signals and the diffuse noise, respectively. $DDR_i = (\Phi_S^{ii}+\Phi_V^{ii})/\Phi_N^{ii}$ represents the true local DDR of the i-th channel microphone signal in a linear scale. According to the study in [8], the DDR ratio can be approximated as

$$\sqrt{\frac{DDR_1}{DDR_1+1}\cdot\frac{DDR_2}{DDR_2+1}} \approx \frac{DDR}{DDR+1}, \quad (6)$$

where DDR can be either DDR_1 or DDR_2.

Meanwhile, the input signal in (3) contains two directional signal components: target speech and interference. Thus, the coherence of the directional signal components can be represented as

$$\Gamma_D(k,l) = \Gamma_S(k,l)\left(\sqrt{\frac{SIR_1}{1+SIR_1}\cdot\frac{SIR_2}{SIR_2+1}}\right) \\ +\Gamma_V(k,l)\left(\sqrt{\frac{1}{SIR_1+1}\cdot\frac{1}{SIR_2+1}}\right), \quad (7)$$

where $\Gamma_S(k,l)$ and $\Gamma_V(k,l)$ are the coherences of the target speech and the directional interference, respectively. The $SIR_i = \Phi_S^{ii}/\Phi_V^{ii}$ represents the target speech of the interference power ratio (SIR) at the i-th channel microphone. In [8], it was shown that the ratio of the SIR has almost the same value in the two channels. Thus, SIR can be approximated as

$$\sqrt{\frac{SIR_1}{SIR_1+1}\cdot\frac{SIR_2}{SIR_2+1}} \approx \frac{SIR}{SIR+1}. \quad (8)$$

By substituting (6), (7), and (8) into (5), the coherence of the noisy observation can be rewritten as

$$\Gamma_Y = \Gamma_S G\cdot K + \Gamma_V(1-G)\cdot K + \Gamma_N(1-K), \quad (9)$$

where $G = SIR/(SIR+1)$ and $K = DDR/(DDR+1)$ are the normalized SIR and DDR, respectively. We note that $G\cdot K$, which represents the normalized speech to interference plus diffuse noise ratio (SINR), is bounded as $0 \leq G\cdot K \leq 1$. At higher SINRs, $G\cdot K$ approaches 1 and, thus, there is a high probability of speech presence and vice versa. Accordingly, $G\cdot K$ can be used as a criterion for determining the *a priori* SAP.

3.2. Proposed *a priori* SAP estimator

In the proposed algorithm, we assume that the target speaker is located in front of a listener surrounded by diffuse noise, and that the interference is coming from an arbitrary direction. The interference can be anywhere around the listener except the frontal direction. In addition, we utilize a real-value coherence function of the diffuse noise field [11]:

$$\hat{\Gamma}_N(k) = sinc\left(\frac{2\pi k f_s d}{N\cdot c}\right), \quad (10)$$

where d is the microphone spacing, N is the maximum frequency bin index, and f_s and $c \approx 340m/s$ represent the sampling frequency and the speed of sound, respectively. The target speech and interference are assumed to be generated from a single well-defined directional sound source, and thus the signals received by the two microphones are perfectly coherent except for a time delay. The coherence function of the directional signal, $\hat{\Gamma}_C$, is then expressed as in [8, 12], as follow:

$$\hat{\Gamma}_C(k) = e^{j2\pi k f_s (d/(N\cdot c))\sin\theta}, \quad (11)$$

where θ is the angle of incidence. Based on the assumption that the target speaker is located in the frontal direction, we approximate the coherence of the target speech to be, $\Gamma_S(k) \approx 1$. Therefore, the noisy coherence function in (9) can be rewritten as

$$\Gamma_Y = GK + (\cos\beta + j\sin\beta)(1-G)K + \hat{\Gamma}_N(1-K), \quad (12)$$

where $\beta = 2\pi k f_s(d/(N\cdot c))\sin\theta_v$ and θ_v is the angle of interference.

First, to obtain the normalized SIR G, we compute the MSC by taking the absolute square of Γ_Y. After rearranging the terms, we can obtain the following quadratic equation:

$$aG^2 + 2bG + c = 0. \tag{13}$$
$$\begin{aligned} a &= 2K^2(1-\cos\beta) \\ b &= K(1-\cos\beta)(\Gamma_N(1-K)-K) \\ c &= \Gamma_N(1-K)(2\cos\beta K + \Gamma_N(1-K)) + K^2 - |\Gamma_Y|^2 \end{aligned}$$

By solving quadratic equation in (13), we obtain the two solutions $G_1 = (-b+\sqrt{b^2-ac})/a$ and $G_2 = (-b-\sqrt{b^2-ac})/a$. It should be noted that the normalized SIR can be either G_1 or G_2, depending on the dominant power of the signal. In a region where the power of the target speech is dominant, G_1 represents the normalized SIR we are seeking. On the other hand, in a region where the power of the interference is dominant, G_2 tracks the normalized SIR. To account for this power dependency, we selectively choose between G_1 and G_2, depending on the dominant power of the current TF-unit:

$$G = G_1 \cdot \gamma + G_2 \cdot (1-\gamma) = \frac{(2\gamma-1)\sqrt{b^2-ac}-b}{a}, \tag{14}$$

where $\gamma = 1$ for a target-dominant region and $\gamma = 0$ for an interference-dominant region. Thus, to determine γ, we must determine whether the processing region is dominated by the target signal or interference. To this end, we utilize the phase information of the noisy input signals. Since it is assumed that the target speaker is located in the frontal direction, any nonzero phase difference between the input signals is caused by interference. Particularly, when the interference power is dominant, the cross PSD Φ_{YY}^{12} will exhibit a nonzero phase. Thus, we determine the classification parameter γ as

$$\gamma = \begin{cases} 1 & \text{if } |\angle\Phi_{YY}^{12}| < \delta \\ 0 & \text{otherwise,} \end{cases} \tag{15}$$

where δ is a phase threshold, which is empirically set to 0.6 in radian. Now, after obtaining $\Gamma_Y(k,l)$ and $\hat{\Gamma}_N(k,l)$ using (4) and (10), we can readily obtain the normalized SIR, G, using (13)-(15), but only if the terms $\cos\beta$ and K are available.

Several algorithms have been proposed to compute the DDR from a noisy speech signal [6, 13]. However, these algorithms were developed under the assumption that there is a single directional target signal in a diffuse noise field. Thus, in cases where the target speech is corrupted by a directional interference plus diffuse noise, previous algorithms may produce biased DDR values. As such, in this paper, we use the minimum PSD tracking method [14] to obtain the normalized DDR. Based on the fact that the minimum power of noisy speech is approximately the same as the power level of the diffuse noise, DDR can be estimated using a pre-measured minimum PSD of the noisy input. To compute the ith-channel local minimum of the noisy speech power, Φ_m^i, we use a non-linear rule with an averaging technique [14]. Then, we can compute the DDR as follows:

$$D\hat{D}R_i = \max\left(\frac{\Phi_{YY}^{ii} - \Phi_m^i}{\Phi_m^i}, 0\right). \tag{16}$$

By substituting (16) into (6), the normalized DDR can be computed.

To obtain $\cos\beta$, we can use the method introduced in [7, 8], where, first, we take the real and imaginary parts of the observation coherence in (12):

$$\begin{aligned} \Re &= GK(1-\cos\beta) + \Gamma_N(1-K) + \cos\beta K, \\ \Im &= \sin\beta(1-G)K. \end{aligned} \tag{17}$$

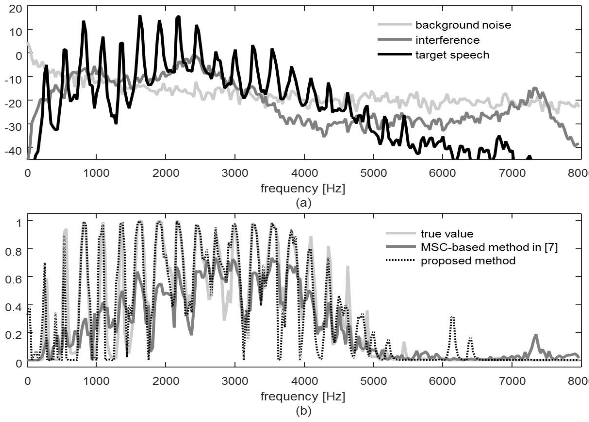

Figure 1: (a) PSDs of the input signal components (b) comparison of the true and estimated SINRs

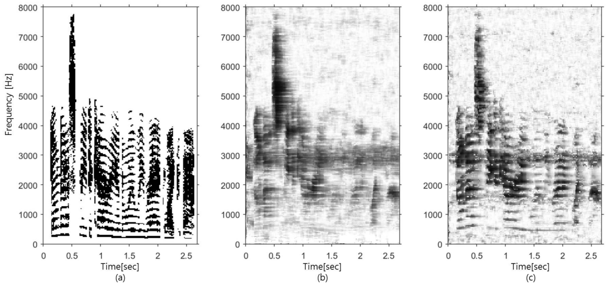

Figure 2: Comparison of (a) the true and estimated *a priori* SAPs using (b) the MSC-based estimator [7] and (c) the proposed estimator (21)

After a few rearrangement steps, the above real and imaginary terms can be combined into a single equation, as follows:

$$(\sin\beta K - \Im)(1-\cos\beta) = \sin\beta(\Re\cos\beta K - \Gamma_N(1-K)). \tag{18}$$

Squaring both sides of (18) and using the fact that $\cos^2\beta + \sin^2\beta = 1$, we have:

$$\cos^2\beta(\Im^2 + (\Re - K - \Gamma_N(1-K))^2) - 2\Im^2\cos\beta + \Im^2 - (\Re - K - \Gamma_N(1-K))^2 = 0. \tag{19}$$

Thus, by solving the quadratic equation, $\cos\beta$ is obtained as

$$\cos\beta = \frac{\Im^2 \pm (\Re - K - \Gamma_N(1-K))^2}{\Im^2 + (\Re - K - \Gamma_N(1-K))^2}. \tag{20}$$

In (20), because the plus sign yields an indeterminate equation that cannot be solved uniquely, we take the minus sign. By substituting (20) into (14), we obtain the normalized SIR G. Finally, we directly approximate the probability of speech absence using a combination of G and K, as given by:

$$\hat{q} = 1 - G \cdot K = \frac{a - K\big((2\gamma-1)\sqrt{b^2-ac}-b\big)}{a}. \tag{21}$$

4. Simulation

Next, we examined the performance of the proposed *a priori* SAP estimator and further compared it with those of existing techniques, including the single-channel SNR-based estimator in [4], the DDR-based estimator in [6], and the MSC-based estimator in [7]. We extracted speech sentences from TIMIT databases for the target signals and speech-like AR random processes for the interferences binaurally convolved with HRIR

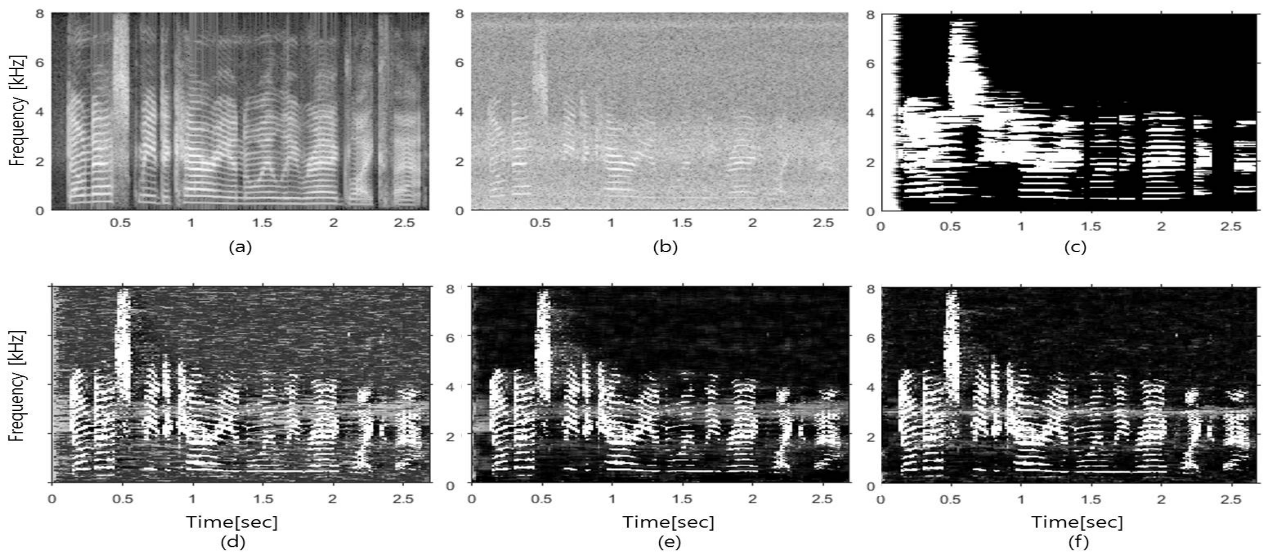

Figure 3: The spectra of (a) clean and (b) noisy speech signals; SPP results obtained using the TF-unit dependent SAP controlled by (c) the SNR-based [4], (d) the DDR-based [6], (e) the MSC-based [7] and (f) the proposed estimators

pairs, corresponding to the target and interference directions, respectively. To simulate speech-like AR random processes, we obtained the AR coefficients from real speech utterances [15]. Later, we added binaural diffuse noise signals based on the SNR. We segmented the noisy input signal into subframes of 512 samples with a 50% overlap using a sine window at a 16 kHz sampling rate. All the implementation parameters for the conventional algorithms were set to the values suggested in the publications. We used 1st-order recursive averaging to estimate the PSD and set the smoothing factor to 0.8 for all algorithms.

Figs.1-3 show the graphical comparisons for a situation in which the frontal target speaker, corrupted by a directional interference at 90° with a 5 dB SIR, was surrounded by modeled diffuse noise at a 5 dB SNR. First, Fig. 1 shows a comparison of the true and estimated combinations of the normalized SIR and DDR. We can see that the proposed method yielded more accurate results than the method used in [7], especially in the region where interference has significant power. Fig. 2 compares the final *a priori* SAP estimates, which also show that the proposed algorithm extracted the target speech with a higher accuracy than the method in [7]. The performance of the method in [7] is seriously degraded by the existence of directional interference. Fig. 3 shows the SPP maskers obtained using the evaluated estimators. We can see from the plots that the SPP estimates based on the proposed *a priori* SAP are more accurate than those by conventional algorithms. In particular, a comparison of the DDR-based (Fig. 3(d)) and the MSC-based (Fig. 3(e)) methods with the proposed technique in Fig. 3(f) shows that the target speech is better extracted from the interference components. We further analyzed the performance of the proposed SPP estimator by calculating the receiver operating characteristic (ROC) curve, which is a parametric plot of the correct detection rate versus the false alarm rate [16]. We used ten pairs of speech sentences and averaged the results obtained for each. We set the direction of the interferences to 90° in samples (a) and (b), and to 230° in samples (c) and (d). We then added binaurally recorded mensa noise from the ETSI database. We set the overall input SIR and DDR to 0 dB and 5 dB, respectively, to generate corresponding SIR and DDR conditions. The obtained ROC curves are illustrated in Figs. 4, which show that the SPP obtained using the proposed *a priori* SAP estimator achieves a significantly better detection rate than do the conventional algorithms for most of SNR and noise conditions.

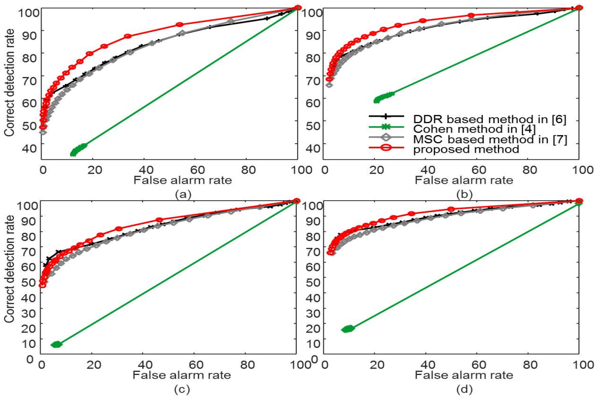

Figure 4: ROC curves for (a) 0 dB, (b) 5 dB SIR and DDR with 90° degree interference and (c) 0 dB and (d) 5 dB SIR and DDR with 230° degree interference

5. Conclusions

In this paper, we addressed the issue of *a priori* SAP estimation for the SPP estimator. We estimated the *a priori* SAP using the combined parameters of normalized SIR and DDR. We derived the normalized SIR using the root of the quadratic equation from the MSC and obtained the DDR from the pre-measured minimum PSD of noisy speech. Unlike conventional methods, we developed the signal model under the assumption of a complex acoustic environment. Thus, the detection accuracy of the proposed estimator can be increased in a mixed sound field. Computer simulations indicate that the proposed method outperforms earlier works with respect to the ROC curve.

6. Acknowledgements

This research was supported by the Ministry of Trade, Industry & Energy(MOTIE), Korea Institute for Advancement of Technology(KIAT) through the Encouragement Program for The Industries of Economic Cooperation Region.

7. References

[1] M. Souden, J. Chen, J. Benesty, and S. Affes, “Gaussian model-based multichannel speech presence probability,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 5, pp. 1072–1077, 2010.

[2] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[3] T. Gerkmann, C. Breithaupt, and R. Martin, “Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 5, pp. 910–919, 2008.

[4] I. Cohen and B. Berdugo, “Speech enhancement for non-stationary noise environments,” *Signal processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

[5] M. Souden, J. Chen, J. Benesty, and S. Affes, “An integrated solution for online multichannel noise tracking and reduction,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 7, pp. 2159–2169, 2011.

[6] M. Taseska and E. A. Habets, “MMSE-based blind source extraction in diffuse noise fields using a complex coherence-based a priori sap estimator,” in *Acoustic Signal Enhancement; Proceedings of IWAENC 2012; International Workshop on*. VDE, 2012, pp. 1–4.

[7] Y. Ji, Y. Baek, and Y.-c. Park, “A priori SAP estimator based on the magnitude square coherence for dual-channel microphone system,” in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4415–4419.

[8] N. Yousefian, P. C. Loizou, and J. H. Hansen, “A coherence-based noise reduction algorithm for binaural hearing aids,” *Speech Communication*, vol. 58, pp. 101–110, 2014.

[9] N. Yousefian and P. C. Loizou, “A dual-microphone speech enhancement algorithm based on the coherence function,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 2, pp. 599–609, 2012.

[10] H. Abutalebi, H. Sheikhzadeh, R. Brennan, and G. Freeman, “A hybrid subband adaptive system for speech enhancement in diffuse noise fields,” *Signal Processing Letters, IEEE*, vol. 11, no. 1, pp. 44 – 47, jan. 2004.

[11] I. McCowan and H. Bourlard, “Microphone array post-filter based on noise field coherence,” *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 6, pp. 709 – 716, nov. 2003.

[12] M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications*. Springer, 2001.

[13] A. Schwarz and W. Kellermann, “Coherent-to-diffuse power ratio estimation for dereverberation,” *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 23, no. 6, pp. 1006–1018, 2015.

[14] S. Rangachari and P. C. Loizou, “A noise-estimation algorithm for highly non-stationary environments,” *Speech communication*, vol. 48, no. 2, pp. 220–231, 2006.

[15] T. Kinnunen, R. Saeidi, F. Sedlák, K. A. Lee, J. Sandberg, M. Hansson-Sandsten, and H. Li, “Low-variance multitaper MFCC features: a case study in robust speaker verification,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 7, pp. 1990–2001, 2012.

[16] H. Momeni, E. A. Habets, and H. R. Abutalebi, “Single-channel speech presence probability estimation using inter-frame and inter-band correlations,” in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 2903–2907.