A PRIORI SAP ESTIMATOR BASED ON THE MAGNITUDE SQUARE COHERENCE FOR DUAL-CHANNEL MICROPHONE SYSTEM

Youna Ji, Yonghyun Baek, Young-cheol Park

Computer and Telecomm. Eng. Division
Yonsei University Wonju, Korea

ABSTRACT

In this paper, we present a time-frequency (TF)-dependent *a priori* speech absence probability (SAP) estimator utilizing the magnitude square coherence (MSC) between two microphone signals. It is shown that the normalized SNR can be numerically computed from the MSC by solving a quadratic equation. Based on the fact that the normalized SNR is bounded between 0 and 1, we directly use it for the probability of speech absence in each TF-unit. Since this approach does not require prior statistical knowledge of noise and speech, it is not affected by the performance of the noise PSD estimator. Furthermore, unlike the conventional SNR-based estimator, additional mapping strategy is unnecessary. The algorithm was evaluated using the receiver operating characteristic (ROC) curve and it attained higher correct detection rate at a given false-alarm rate than the conventional algorithms.

Index Terms— speech presence probability, speech absence probability, magnitude square coherence

1. INTRODUCTION

The speech presence probability (SPP) estimator plays an important role in the performance of the speech enhancement system. A general SPP estimator can be derived under the assumption that the spectral coefficients of speech and noise can be modeled as complex Gaussian random variables [1]. *A posteriori* SPP is computed per time-frequency (TF) unit in the short time Fourier transform (STFT)-domain based *a priori* and *a posteriori* SNRs and *a priori* speech absence probability (SAP).

Since theoretically, *a priori* SAP does not depend on the observation, it can be set as a fixed value [1, 2, 3]. The estimator presented in [2] is an example of the single-channel SPP estimator employing a fixed *a priori* SNR and SAP. However, in practice, it can be assumed that the SAP is varying with time and frequency, depending on the words spoken [3]. Hence, it is more appropriate to estimate the *a priori* SAP in each TF-unit instead of using a fixed value.

Several algorithms have been proposed to estimate and update *a priori* SAP. The well-known single-channel soft decision approach was established in [4], where it was used that the neighboring frequency bins of consecutive frames in the speech presence region have high correlation. A similar approach was applied to the multichannel system in [5], by taking into account the local and global variations of SNR. In these approaches, the estimated SNRs were averaged and mapped onto a value between zero to one to use it as an estimate of SAP. However, although SNR is a strongly correlated with SAP, the accuracy of the obtained SAP is highly affected by the noise estimation performance and the mapping function. Recently, another multichannel *a priori* SAP estimator was proposed in [6] where an estimate of the direct to diffuse ratio (DDR) was utilized. Since this estimator does not require statistical information of noise or speech, the detection accuracy is decoupled with the noise estimation performance. However, it still requires a mapping function to obtain an SPP estimate.

In this paper, we propose *a priori* SAP estimator based on the magnitude square coherence (MSC) of the dual-channel microphone signals. We show that the normalized SNR can be obtained from MSC by solving a quadratic equation. Then, the *a priori* SAP is obtained directly from the estimated SNR without an additional mapping process. As a result, the accuracy of the proposed *a priori* SAP estimator is independent of the performance of the noise PSD estimator.

2. GENERAL SPEECH PRESENCE PROBABILITY IN THE DUAL-CHANNEL SYSTEM

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

$$Y_m(k,l) = X_m(k,l) + N_m(k,l), m = 1, 2, \qquad (1)$$

where $X_m(k,l) = S(k,l)A_m(k,l)$, $S(k,l)$ is the target speech source, $A_m(k,l)$ is the acoustic path from the speech source to the mth-channel microphone. k and l denote the frequency bin and frame indices, respectively. $N_m(k,l)$ are assumed to be diffuse noise propagating in all directions simultaneously with equal power and random phase [7, 8]. 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 PSD matrix of $\mathbf{y}(k,l)$ is defined as $\mathbf{\Phi}_{yy}(k,l) = E[\mathbf{y}(k,l)\mathbf{y}^H(k,l)]$.

Lets assume that $H_1(k,l)$ and $H_0(k,l)$ are two-states hypotheses which represent speech presence and absence, respectively. Then, under the assumption that the desired speech and noise components can be modelled as complex multivariate Gaussian random variables, the multichannel *a posteriori* SPP estimate is obtained as [1]

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

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 *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 SPEECH ABSENCE PROBABILITY ESTIMATOR

3.1. Normalized SNR and SAP

In a recent study [9], a coherence-based noise reduction technique was proposed in a situation that a frontal target speaker was present together with an undesired interference. The technique utilized the real and imaginary parts of the coherence function between the input signals as a criterion for estimating the normalized SNR.

The coherence between two microphone signals can be calculated as

$$\Gamma_Y(k,l) = \frac{\phi_{YY}^{12}(k,l)}{\sqrt{\phi_{YY}^{11}(k,l)\phi_{YY}^{22}(k,l)}}, \quad (3)$$

where $\phi_{YY}^{ij} = E[Y_i(k,l)Y_j^*(k,l)], i,j = 1,2$ are cross- and auto-PSDs of the microphone signals. In [10], it was shown that the coherence of dual-channel observations can be expressed as a weighted sum of speech and noise coherences, as given by

$$\begin{aligned} \Gamma_Y(k,l) = \Gamma_X(k,l)&\left(\sqrt{\frac{SNR_1}{1+SNR_1}\frac{SNR_2}{1+SNR_2}}\right) \\ &+\Gamma_N(k,l)\left(\sqrt{\frac{1}{1+SNR_1}\frac{1}{1+SNR_2}}\right), \end{aligned} \quad (4)$$

where SNR_1 and SNR_2 respectively represent the true local SNR of first and second microphone signals in linear scale. Also, it was shown in [9], the SNR ratio is relatively independent of the target direction, i.e., $SNR_1/(1+SNR_1) \approx SNR_2/(1+SNR_2)$. Therefore, we define the normalized SNR as

$$G = \sqrt{\frac{SNR_1}{1+SNR_1}\frac{SNR_2}{1+SNR_2}} \approx \frac{SNR}{1+SNR}, \quad (5)$$

where SNR can be either SNR_1 or SNR_2. It should be noted that the normalized SNR is bounded as $0 \le G \le 1$. At higher SNR, we have $G \approx 1$, and thus there is a high probability of speech presence and vice versa. Thus, the normalized SNR G is strongly correlated to the probability of speech absence. The variable G can be computed using the method suggested in [9]. However, it was developed under the assumption that a target speaker is located in the frontal direction, which is impractical in real environments. It is also possible to modify the method in [9] for a target speaker in an arbitrary direction. But experimental results revealed that the accuracy of the estimated SNR is still problematic for non-frontal target speakers. Thus, in this paper, we propose a method of computing the normalized SNR using the MSC, in order to achieve accurate estimates of the normalized SNR even for the target speaker in an arbitrary direction.

By assuming a diffuse noise field, the noise coherence, $\Gamma_N(k,l)$ in (4) can be replaced with a real-valued analytical model [11]:

$$\hat{\Gamma}_N(k) = sinc\left(\frac{2\pi k f_s d}{K\cdot c}\right), \quad (6)$$

where d is the microphone spacing, K is the maximum frequency bin index, f_s and $c \approx 340m/s$ represent the sampling frequency and the speed of sound, respectively. On the other hand, the speech coherence function due to the speech signal incident from angle θ can be calculated as [12]

$$\hat{\Gamma}_X(k) = e^{j2\pi k f_s (d/(K\cdot c))\sin\theta}. \quad (7)$$

Thus by substituting (6) and (7) into (4), the observation coherence function can be rewritten as,

$$\Gamma_Y = [\cos(\alpha) + j\sin(\alpha)]G + \hat{\Gamma}_N(1-G), \quad (8)$$

where $\alpha = 2\pi k f_s(d/(K\cdot c))\sin\theta$. For the sake of simplicity, we omit frequency and frame indices.

By taking the absolute square of Γ_Y, we can obtain the MSC:

$$\Psi_Y = |\Gamma_Y|^2 = aG^2 - 2bG + \hat{\Gamma}_N^2, \quad (9)$$

where $a = 1 - 2\hat{\Gamma}_N\cos\alpha + |\hat{\Gamma}_N|^2$ and $b = \hat{\Gamma}_N(\hat{\Gamma}_N - \cos\alpha)$. Now, the MSC can be rearranged into a quadratic equation:

$$aG^2 - 2bG + \hat{\Gamma}_N^2 - \Psi_Y = 0. \quad (10)$$

The normalized SNR G can be computed by solving the quadratic equation in (10),

$$G = \frac{b \pm \sqrt{\hat{\Gamma}_N^2(\cos^2\alpha - 1) + a\Psi_Y}}{a}. \quad (11)$$

Since Γ_Y and $\hat{\Gamma}_N$ can be computed using equations (3) and (6), the normalized SNR G is readily obtained, only if the term $\cos\alpha$ is available. To obtain $\cos\alpha$, the real and imaginary parts of the observation coherence in (8) are taken:

$$\begin{aligned} \Re &= \cos\alpha G + \hat{\Gamma}_N(1-G), \\ \Im &= \sin\alpha G. \end{aligned} \quad (12)$$

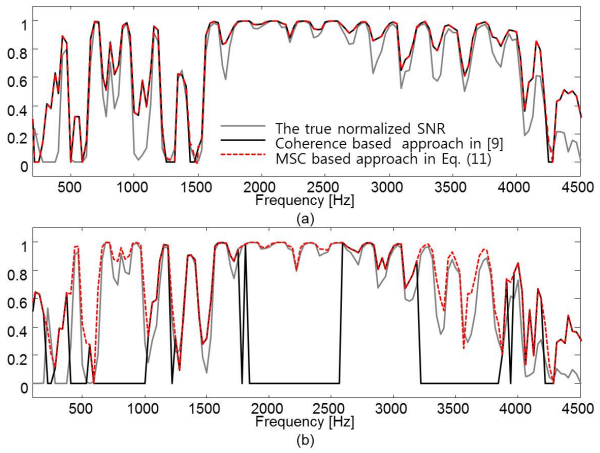

Fig. 1: The true and estimated normalized SNRs at (a) 0° and (b) 90°.

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

$$\Im(\cos\alpha - \hat{\Gamma}_N) = \sin\alpha(\Re - \hat{\Gamma}_N). \tag{13}$$

Squaring both sides of (13) and using the fact that $\sin^2\alpha = 1 - \cos^2\alpha$, we can have

$$\begin{aligned}(\Im^2 + (\Re - \hat{\Gamma}_N)^2)\cos^2\alpha - \\ 2\hat{\Gamma}_N\Im^2\cos\alpha + \Im^2\hat{\Gamma}_N^2 - (\Re - \hat{\Gamma}_N)^2 = 0.\end{aligned} \tag{14}$$

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

$$\cos\alpha = \frac{\hat{\Gamma}_N\Im^2 \pm \sqrt{\nu}}{\Im^2 + (\Re - \hat{\Gamma}_N)^2}, \tag{15}$$

where $\nu = (\Re - \hat{\Gamma}_N)^2(\Im^2(1 - \hat{\Gamma}_N^2) + (\Re - \hat{\Gamma}_N)^2)$. Using the fact that $\cos\alpha$ should be 1 for the target at 0°, plus sign can be taken as the correct solution. Now substituting (15) into (11), the normalized SNR G can be obtained. In previous studies [4, 5, 13], a mapping function was introduced to convert the estimated SNR to the bounded. However, the normalized SNR G is bounded between 0 and 1, and thus *a priori* SAP can be directly approximated using G without additional mapping function as

$$\begin{aligned}\hat{q} &= 1 - G, \\ &= \frac{1 - \hat{\Gamma}_N\cos\alpha - \sqrt{\hat{\Gamma}_N^2(\cos^2\alpha - 1) + a\Psi_Y}}{a}.\end{aligned} \tag{16}$$

The accuracy of the proposed MSC-based SNR estimator is evaluated and compared with that of the method in [9]. Fig. 1 compares the obtained SNRs for the cases of frontal ($\theta = 0°$) and non-frontal ($\theta = 90°$) target speakers, respectively. It should be mentioned that the method in [9] was modified to suit it for target speakers in an arbitrary direction. The results show that the proposed MSC-based SNR estimator provides robust results even for the non-frontal target speaker unlike the conventional method in [9]

3.2. Modification for Practical Consideration

In many of the previous studies [2, 4, 5], it was shown that the combination of local and global variables, could provide performance improvement over the use of a single variable. Thus for the proposed SAP estimator, we combine the local and global estimates of the *a priori* SAP via a multiplicative combination, as given by

$$\hat{q}(k,l) = \hat{q}_{\texttt{local}}(k,l) \cdot \hat{q}_{\texttt{glob}}(k,l). \tag{17}$$

The difference between $\hat{q}_{\texttt{local}}(k,l)$ and $\hat{q}_{\texttt{glob}}(k,l)$ is the number of frequency bins over which the observation coherence $\Gamma_Y(k,l)$ is averaged using

$$\bar{\Gamma}_Y(k,l) = \frac{1}{2\cdot k_d + 1}\sum_{k=k-k_d}^{k=k+k_d}\Gamma_Y(k,l). \tag{18}$$

The averaging over adjacent frequency bins results in a reduction of random fluctuations. The range of averaging is a trade-off between low-variance and fine-structure. In this paper, $k_d = 1$ and $k_d = 10$ were used to estimate the local and global variables, $\hat{q}_{\texttt{local}}(k,l)$ and $\hat{q}_{\texttt{global}}(k,l)$, respectively. Fig. 2 depicts the (a) local, (b) global SAPs and (c) the combined-SAP using (16) for a 0 dB noisy speech signal. It can be seen that the local SAP has high variance but preserves the fine structure of the speech spectrum. On the other hand, the global SAP reduces the variance but has a low degree of precision. The combination of the two positively combines the features of the local and global SAP.

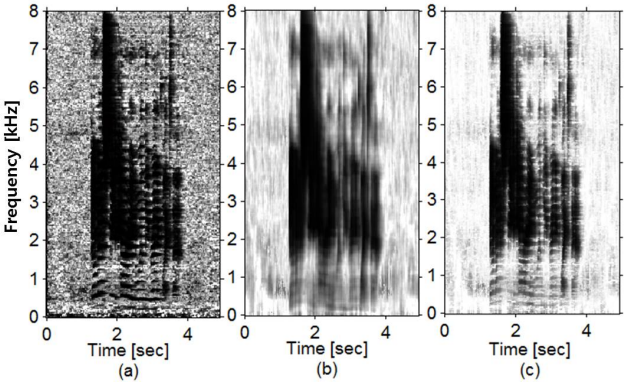

Fig. 2: The estimated (a) local SAP, (b) global SAP and (c) the combined SAP for a 0 dB noisy speech.

4. SIMULATION

Speech sentences in TIMIT databases were extracted and binaurally convolved with HRIR pairs corresponding to the target directions. Later, binaural noise signals taken from ETSI database were added according to SNR. The noisy input signal was segmented into subframes of 512 samples with 50% overlap using a sine window at a 16 kHz sampling rate. The

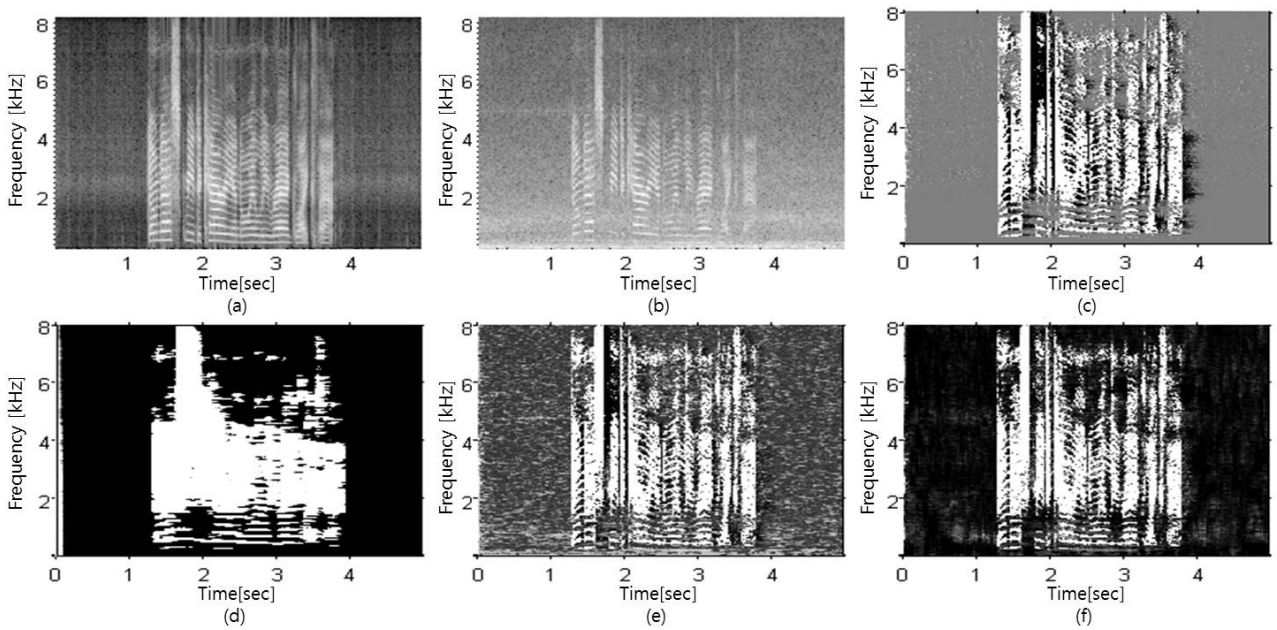

Fig. 3: The spectra of (a) the clean, (b) noisy speech signals, SPP results obtained using (c) a fixed SAP approach [1], and the TF-unit dependent SAP controlled by (d) the SNR-based [4], (e) the DDR-based [6] and (f) the proposed estimator in Eq. (16)

performance of the proposed SAP estimator was compared with those of the single-channel SNR-based estimator in [4], the dual-channel a fixed SAP based approach in [1] and DDR-based estimator in [6]. All implementation parameters for the conventional algorithms were set to the values suggested in the publications. The smoothing factor for the PSD estimation was the same for all algorithms, which was $\alpha_y = 0.7$.

First, Fig. 3 shows the SPP maskers obtained using the evaluated estimators for a target speaker in the frontal direction. Spectrograms of the clean and 0 dB noisy input signals are shown in Fig. 3(a) and (b), respectively. The SPP maskers obtained using the fixed, the SNR- and DDR-based SAPs are shown in Fig. 3(c), (d) and (e), respectively. The fixed SAP approach produces biased SPPs in noise only spectral regions. The SNR-based estimator, on the other hand, clearly distinguishes the speech presence region from the noise regions, but it is difficult to find harmonic structure especially in high SNR regions. The DDR-based estimator provides the detailed harmonic structure of the speech signal, but it also produces a biased SPPs with high variance in noise spectral regions. On the contrary, SPP obtained using the proposed estimator in Fig. 3(f) provides not only fine harmonic structures but also almost unbiased SPP in the noise spectral regions.

The performance of the proposed SPP estimator was further analyzed by calculating the receiver operating characteristic (ROC) curve which is a parametric plot of the correct detection rate versus the false alarm rate [14]. 5 pairs of speech sentences were used and the results obtained for each sentence were averaged. Pink noise and diffuse noise were added to make 0 dB and 5 dB SNR conditions. The obtained ROC curves are depicted in Fig. 4. It can be seen that the SPP obtained using the proposed *a priori* SAP estimator achieves an always significantly higher correct detection rate than the conventional algorithms for all SNR and noise conditions.

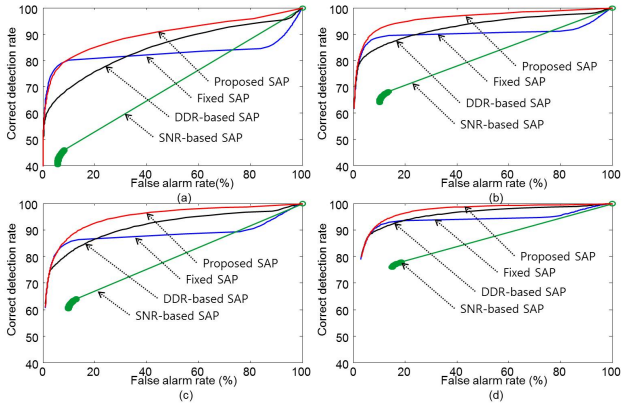

Fig. 4: ROC curves for (a) 0 dB, (b) 5 dB SNR with pink noise, and (c) 0 dB and (d) 5 dB SNR with mensa noise from ETSI database

5. CONCLUSION

In this paper, a new *a priori* SAP estimator was proposed based on the MSC of two microphone signals. The proposed algorithm requires neither an additional mapping function nor prior knowledge of noise or speech statistics. The simulation results showed that the proposed SPP estimator can maintain the harmonic structure of the speech signal and can obtain higher speech detection rate for a given false alarm probability than the conventional algorithms.

6. REFERENCES

[1] Mehrez Souden, Jingdong Chen, Jacob Benesty, and Sofiène 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] Timo Gerkmann, Colin Breithaupt, and Rainer 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.

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

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

[5] Mehrez Souden, Jingdong Chen, Jacob Benesty, and Sofiene 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] Maja Taseska and Emanuël AP 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] H.R. Abutalebi, H. Sheikhzadeh, R.L. Brennan, and G.H. 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.

[8] I.A. 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.

[9] Nima Yousefian, Philipos C Loizou, and John HL Hansen, "A coherence-based noise reduction algorithm for binaural hearing aids," *Speech Communication*, vol. 58, pp. 101–110, 2014.

[10] Nima Yousefian and Philipos 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.

[11] Iain A McCowan and Hervé Bourlard, "Microphone array post-filter based on noise field coherence," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 6, pp. 709–716, 2003.

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

[13] Zhong-Hua Fu and Jhing-Fa Wang, "Speech presence probability estimation based on integrated time-frequency minimum tracking for speech enhancement in adverse environments," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 4258–4261.

[14] Hajar Momeni, Emanuël AP Habets, and Hamid Reza 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.