A Second-Order-Statistics-based Solution for Online Multichannel Noise Tracking and Reduction

Mehrez Souden, Jingdong Chen, Jacob Benesty, and Sofiène Affes

***Abstract*— We propose a second-order-statistics-based approach to online multichannel noise tracking and reduction. We combine the multichannel speech presence probability (MC-SPP) that we proposed in [1] with an alternative formulation of the minima-controlled recursive averaging (MCRA) technique that we generalize from the single- to the multichannel case. Then, we demonstrate the effectiveness of the proposed MC-SPP and multichannel noise estimator by integrating them into variants of the multichannel noise-reduction Wiener filter.**

***Index Terms*— Microphone array, noise estimation, multichannel speech presence probability (MC-SPP), multichannel noise reduction, minima controlled recursive averaging (MCRA).**

I. Introduction

Speech signals are inherently sparse in the time and frequency domains, thereby allowing for continuous tracking and reduction of background noise in speech acquisition systems. Indeed, spotting time instants and frequency bins without/with active speech components is extremely important to update/hold the noise statistics that are needed in the design of noise-reduction filters. When multiple microphones are utilized, the extra space dimension has to be optimally exploited for this purpose.

In [2], the minimum variance distortionless response (MVDR), in particular, and parameterized multichannel Wiener filter (PMWF), in general, were formulated such that they only depend on the noise and noisy data power spectrum density (PSD) matrices when only noise reduction is of interest. Therefore, what one actually needs when implementing these filters are accurate estimates of the noise and noisy data PSD matrices. This can be viewed as a natural extension from the single to the multichannel case. Following the single-channel noise reduction legacy, it seems natural to also generalize the concepts of speech presence probability (SPP) estimation and noise tracking to the multichannel case in order to implement the multichannel noise reduction filters. Recently, the MC-SPP has been theoretically formulated and its advantages were discussed in [1]. In this paper, we first propose a practical implementation of the MC-SPP. Furthermore, an online estimator of the noise PSD matrix which generalizes the MCRA to the multichannel case is provided. Similar to the single-channel scenario, we show how the noise estimation is performed during speech absence only. After investigating the accuracy of the speech detection when multiple microphones are utilized, we combine the multichannel noise estimator with PMWF-based noise reduction methods. The overall proposed scheme performs very well in various conditions: stationary or non-stationary noise in anechoic or reverberant acoustic rooms.

II. Problem Statement

Let $S(k,l)$ be a speech signal impinging on an array of N microphones with an arbitrary geometry. k and l respectively denote the frequency and time frame indices (in the STFT domain). The resulting observations are given by

$$Y_n(k,l) \quad = \quad X_n(k,l) + V_n(k,l), \ n = 1, 2, ..., N, \quad (1)$$

where $X_n(k,l) = G_n(k)S(k,l)$, $G_n(k)$ is the transfer function of the propagation channel between the source and the nth microphone location. $k = 0, ..., K-1$ (K is the STFT length). With this model, the objective of noise reduction is to estimate one of the N clean speech spectra $X_n(k,l)$, $n = 1, 2, ..., N$. Without loss of generality, we choose to estimate $X_1(k,l)$. We define $\mathbf{y}(k,l) \triangleq [Y_1(k,l) \ \cdots \ Y_N(k,l)]^T$.

III. Multichannel Wiener Filter-Based Noise Reduction

It is important to emphasize that our purpose here is to reduce the additive noise the best way we can with no attempt of dereverberation. This has been the objective of numerous research efforts using single or multiple microphones [3], [4], [5], [6], [7], [8]. Nevertheless, while most effective single channel-based processing approaches take advantage of the noise and noisy-data PSD matrices, several multichannel noise reduction techniques require the estimation of the steering vector as a preprocessing stage [8]. It turns out that only the noise and noisy-data PSD matrices are required to reduce the additive noise as in the single-channel case. The PMWF, in general, and MVDR (equivalently its GSC implementation), in particular, are good examples. Indeed, we have shown in [2] that the PMWF is given by

$$\mathbf{h}_{\mathrm{W},\beta}(k,l) = \frac{\mathbf{\Phi}_{vv}^{-1}(k,l)\mathbf{\Phi}_{xx}(k,l)\mathbf{u}_1}{\mu + \xi(k,l)} \quad (2)$$

where $\xi(k,l) = \mathrm{tr}\left\{\mathbf{\Phi}_{vv}^{-1}(k,l)\mathbf{\Phi}_{xx}(k,l)\right\}$,

$$\mu \leq \frac{\tilde{\sigma}}{1-\tilde{\sigma}(k,l)}\xi(k,l), \quad (3)$$

$\tilde{\sigma}(k,l) = \sigma\phi_{x_1x_1}^{-1/2}(k,l)$, and σ is the maximum speech distortion. Note that taking the upper bound in (3) results in maximum noise reduction and a signal distortion of σ. Also, it is straightforward to see from (3) that by imposing no signal distortion ($\sigma = 0$), we obtain the MVDR expression as a particular case of (2) with $\mu = 0$. In order to implement the PMWF, the noise and noisy data PSD matrices have to be properly estimated; this is the purpose of the following section.

IV. Second-Order-Statistics Estimation

Here, our aim is to propose a solution to estimate the PSD matrices of the noise and noise-free data. These matrices are directly involved in the expression of the PMWF-based filters as shown above. We denote the noise and noisy data PSD matrices as $\mathbf{\Phi}_{vv}(k) \triangleq E\left\{\mathbf{v}(k,l)\mathbf{v}^H(k,l)\right\}$ and $\mathbf{\Phi}_{yy}(k) \triangleq E\left\{\mathbf{y}(k,l)\mathbf{y}^H(k,l)\right\}$, respectively. In practice, a first order recursive time-smoothing is used to estimate these PSD matrices from the available data samples. In other words, at time frame l, the estimates of the noisy data statistics are updated recursively [we use the notation $\hat{(\cdot)}$ to denote "the estimate of"]

$$\hat{\mathbf{\Phi}}_{yy}(k,l) = \alpha_y(k,l)\hat{\mathbf{\Phi}}_{yy}(k,l-1)+[1-\alpha_y(k,l)]\,\mathbf{y}(k,l)\mathbf{y}^H(k,l) \quad (4)$$

where $0 \leq \alpha_y(k,l) \leq 1$. As for the noise PSD matrix estimation, we state that any of the known single channel noise estimation methods (e.g., minimum-statistics [9], MCRA [3], [10]) can be extended to the multichannel case. Without loss of generality, we consider a framework that is similar to the one proposed in [3], [10]. More specifically, we recursively estimate the noise statistics as

$$\hat{\mathbf{\Phi}}_{vv}(k,l) = \tilde{\alpha}_v(k,l)\hat{\mathbf{\Phi}}_{vv}(k,l-1)+[1-\tilde{\alpha}_v(k,l)]\,\mathbf{y}(k,l)\mathbf{y}^H(k,l), \quad (5)$$

where $0 \leq \tilde{\alpha}_v(k,l) \leq 1$ and should be small enough when the speech is absent so that $\hat{\mathbf{\Phi}}_{vv}(k,l)$ can follow the noise changes. But when the speech is present, this parameter should be sufficiently large to avoid noise PSD matrix overestimation and speech cancelation. Clearly, the parameter $\tilde{\alpha}_v(k,l)$ is closely related to the detection of speech presence/absence. Similar to the single-channel MCRA, we demonstrate that the MC-SPP, denoted as $p(k,l)$, is directly related to $\tilde{\alpha}_v(k,l)$ as

$$\tilde{\alpha}_v(k,l) = \alpha_v + (1-\alpha_v)p(k,l) \quad (6)$$

where $0 \leq \alpha_v(k,l) \leq 1$.

V. Multichannel Speech Presence Probability

The SPP in the single-channel case has been exhaustively studied [10], [11], [4]. In the multichannel case, the two-state model of speech presence/absence holds as in the single-channel case. In other words, we have

1) $H_0(k,l)$: in which case the speech is absent, i.e.,

$$\mathbf{y}(k,l) = \mathbf{v}(k,l). \quad (7)$$

2) $H_1(k,l)$: in which case the speech is present, i.e.,

$$\mathbf{y}(k,l) = \mathbf{x}(k,l) + \mathbf{v}(k,l). \quad (8)$$

A first attempt to generalize the concept of SPP to the multichannel case was made in [12] where some restrictive assumptions (uniform linear microphone array, anechoic propagation environment, additive white Gaussian noise) were made to develop an MC-SPP. Recently, we have generalized this study and shown that this probability is in the following form [1]

$$p(k,l) = \left\{1 + \frac{q(k,l)}{1-q(k,l)}\,[1+\xi(k,l)]\exp\left[-\frac{\beta(k,l)}{1+\xi(k,l)}\right]\right\}^{-1} \quad (9)$$

where $\xi(k,l)$ is defined in Section III and can be identified as the multichannel *a priori* SNR [1]. Moreover, we have

$$\beta(k,l) \triangleq \mathbf{y}^H(k,l)\mathbf{\Phi}_{vv}^{-1}(k,l)\mathbf{\Phi}_{xx}(k,l)\mathbf{\Phi}_{vv}^{-1}(k,l)\mathbf{y}(k,l), \quad (10)$$

and $q(k,l)$ is the *a priori* SAP. The result in (9)–(10) describes how the multiple microphones' observations can be combined in order to achieve optimal speech detection. It can be viewed as a straightforward generalization of the single-channel SPP to the multichannel case.

A. Estimation of the A Priori Speech Absence Probability

We see from (9) that the *a priori* SAP, $q(k,l)$, needs to be estimated. In single-channel approaches, this probability is often set to a fixed value [4], [6]. However, speech signals are inherently non-stationary. Hence, choosing a time- and frequency-dependent *a priori* SAP can lead to more accurate detectors. Notable contributions that have recently been proposed include [3] where the *a priori* SAP is estimated using a soft decision that takes advantage of the correlation of the speech presence in neighboring frequency bins of consecutive frames. In [10], a single-channel estimator of the *a priori* SAP which is based on minimum statistics tracking was proposed. The method is inspired from [9], but further uses time and frequency smoothing.

In contrast to previous contributions, we propose to use multiple observations captured by an array of microphones to achieve more accuracy in estimating the *a priori* SAP. Theoretically, any of the aforementioned principles (fixed SAP, minimum-statistics, or correlation of the speech presence in neighboring frequency bins of consecutive frames) can be extended to the multichannel case. Without loss of generality, we consider a framework that is similar to the one proposed in [3] and use both long-term and instantaneous variations of the overall observations' energy (with respect to the best estimate of the noise energy). Our method is based on the multivariate statistical analysis [13] and jointly processes the N microphone observations for optimal *a priori* SAP estimation.

We define the following two terms

$$\psi(k,l) \triangleq \mathbf{y}^H(k,l)\hat{\mathbf{\Phi}}_{vv}^{-1}(k,l)\mathbf{y}(k,l), \quad (11)$$

$$\tilde{\psi}(k,l) \triangleq \mathrm{tr}\left[\hat{\mathbf{\Phi}}_{vv}^{-1}(k,l)\hat{\mathbf{\Phi}}_{yy}(k,l)\right]. \quad (12)$$

Both terms will be used for *a priori* SAP estimation. Note first that in the particular case $N=1$, $\tilde{\psi}(k,l)$ boils down to the well known *a posteriori* SNR [3], [10], [9] in the single-channel case. Besides, $\psi(k,l)$ is nothing but the instantaneous version of $\tilde{\psi}(k,l)$. We have $\tilde{\psi}(k,l) \geq N$ and large values of $\psi(k,l)$ and $\tilde{\psi}(k,l)$ would indicate the speech presence, while small values (close to N) would indicate speech absence. By analogy to the single channel-case, $\psi(k,l)$ and $\tilde{\psi}(k,l)$ can be identified as the instantaneous and long-term estimates of the multichannel *a posteriori* SNR, respectively. Consequently, considering both terms in (11) and (12) to have a prior estimate of the SAP amounts to assessing the instantaneous and long-term averaged observations' energies compared to the best available noise statistics estimates and deciding whether the speech is *a priori* absent or present.

Now, we see from the definitions in (11) and (12) that in order to control the false alarm rate, two thresholds ψ_0 and $\tilde{\psi}_0$ have to be chosen such that

$$\begin{aligned} Prob\left[\psi(k,l) \geq \psi_0 | H_0(k,l)\right] &\leq \epsilon, \\ Prob\left[\tilde{\psi}(k,l) \geq \tilde{\psi}_0 | H_0(k,l)\right] &\leq \epsilon, \end{aligned} \tag{13}$$

where ϵ denotes a certain significance level that we choose as $\epsilon = 0.01$ [3]. In theory, the distributions of $\psi(k,l)$ and $\tilde{\psi}(k,l)$ are required to determine ψ_0 and $\tilde{\psi}_0$. In practice, it is very difficult to determine the two probability density functions (PDFs). To circumvent this problem, we make the following two assumptions for *noise only frames*.

- *Assumption 1*: the vectors $\mathbf{y}(k,l)$ are Gaussian, independent, and identically distributed with mean $\mathbf{0}$ and covariance $\mathbf{\Phi}_{vv}(k,l)$.
- *Assumption 2*: the noise PSD matrix can be approximated as a sample average of L periodograms (we further assume that these periodograms are independent for ease of analysis), i.e.,

$$\hat{\mathbf{\Phi}}_{vv}(k,l) \approx \frac{1}{L}\sum_{i=1}^{L} \mathbf{y}(k,l_i)\mathbf{y}^H(k,l_i) \tag{14}$$

  where l_i is a certain time index of a speech-free frame preceding the lth one. Following this assumption, $\hat{\mathbf{\Phi}}_{vv}(k,l)$ has a complex Wishart distribution $W_N\left(\mathbf{\Phi}_{vv}(k,l), L\right)$ [in the following, we will use the notation $\hat{\mathbf{\Phi}}_{vv}(k,l) \sim W_N\left(\mathbf{\Phi}_{vv}(k,l), L\right)$] [13].

Using *Assumption 1* and *Assumption 2*, we find that $\psi(k,l)$ has a Hotelling's T^2 distribution with PDF and cumulative distribution function (CDF) respectively expressed as [13]

$$f_{\psi}(x) = \frac{\Gamma(L+1)}{L\Gamma(N)\Gamma(L-N+1)} \frac{\left(\frac{x}{L}\right)^{N-1}}{\left(1+\frac{x}{L}\right)^{L+1}} u(x) \tag{15}$$

$$\begin{aligned} \mathcal{F}_{\psi}(x) = &\left(\frac{x}{L}\right)^{N} \frac{L\Gamma(L)}{\Gamma(N+1)\Gamma(L-N+1)} \times \\ &{}_2F_1\left(N, L+1; N+1; -\frac{x}{L}\right) u(x) \end{aligned} \tag{16}$$

where ${}_2F_1\left(\cdot,\cdot;\cdot;\cdot\right)$ is the hypergeometric function [13], [14], and $u(x) = 1$ if $x \geq 1$ and 0 otherwise.

Now, we turn to the estimation of $\tilde{\psi}_0$. To this end, we use *Assumption 1* and further suppose that, similar to $\hat{\mathbf{\Phi}}_{vv}(k,l)$, $\hat{\mathbf{\Phi}}_{yy}(k,l)$ can be approximated by a sample average of L periodograms. In order to determine the PDF of $\tilde{\psi}(k,l)$, we use the fact that for two random $d \times d$–dimensional matrices $\mathbf{H} \sim W_d(\mathbf{\Sigma}, m_H)$ and $\mathbf{E} \sim W_d(\mathbf{\Sigma}, m_E)$, the distribution of $\mathrm{tr}\left\{\mathbf{H}\mathbf{E}^{-1}\right\}$ can be approximated by cF where $F \sim F_{a,b}$ (F distribution with a and b degrees of freedom) where [13], [15]

$$a = dm_H,\ b = 4 + \frac{a+2}{B-1},\ c = \frac{a(b-2)}{b(m_E-d-1)}$$

$$B = \frac{(m_E+m_H-d-1)(m_E-1)}{(m_E-d-3)(m_E-d)}.$$

Specifically, the PDF and CDF corresponding to $F_{a,b}$ are [13]

$$f_{\tilde{\psi}}(x) = \frac{\sqrt{\frac{(ax)^a b^b}{(ax+b)^{a+b}}}}{x\mathcal{B}\left(\frac{a}{2},\frac{b}{2}\right)} u(x) \tag{17}$$

$$\mathcal{F}_{\tilde{\psi}}(x) = I_{\frac{ax}{ax+b}}\left(\frac{a}{2},\frac{b}{2}\right) u(x). \tag{18}$$

This approximation is valid for real matrices and we found that it gives good results in the investigated scenario for $\tilde{\psi}(k,l)$ [i.e., replacing $\mathbf{H}$ and $\mathbf{E}$ by $\hat{\mathbf{\Phi}}_{yy}(k,l)$ and $\hat{\mathbf{\Phi}}_{vv}(k,l)$, respectively] by choosing $m_E = m_H = L$ and $d = 2N$. Note again that we are assuming that $\hat{\mathbf{\Phi}}_{yy}(k,l)$ and $\hat{\mathbf{\Phi}}_{vv}(k,l)$ have the same mean since we are considering *noise only frames*.

Once we determine ψ_0 and $\tilde{\psi}_0$ using (13) jointly with (16) and (18), we have to take into account the variations of both $\psi(k,l)$ and $\tilde{\psi}(k,l)$ in order to devise an accurate estimator of the *a priori* SAP. Hence, we propose a procedure which is inspired from the work of Cohen in [3], [10]. We first propose the following three estimators $\hat{q}_{\mathrm{local}}(k,l)$, $\hat{q}_{\mathrm{global}}(k,l)$, and $\hat{q}_{\mathrm{frame}}(l)$ which are described in the following.

For a given frequency bin, we estimate the local (at frequency bin k) *a priori* SAP as [3]

$$\hat{q}_{\mathrm{local}}(k,l) = \begin{cases} 1 & \text{if } \tilde{\psi}(k,l) < N \\ & \quad \text{and } \psi(k,l) < \psi_0 \\ \frac{\tilde{\psi}_0 - \tilde{\psi}(k,l)}{\tilde{\psi}_0 - N} & \text{if } N \leq \tilde{\psi}(k,l) < \tilde{\psi}_0 \\ & \quad \text{and } \psi(k,l) < \psi_0 \\ 0 & \text{else.} \end{cases}$$

When $\psi(k,l)$ and $\tilde{\psi}(k,l)$ are sufficiently large, it is assumed that the speech is *a priori* locally present. If $\psi(k,l)$ is lower than ψ_0 and $\tilde{\psi}(k,l)$ is lower than its minimum theoretical lower value N, we decide that the speech is *a priori* absent. In mild situations, a soft transition from speech to non-speech decision is performed.

Note that the condition on $\psi(k,l)$ in (19) represents a local decision that the speech is assumed to be *a priori* absent or present using the information retrieved from a single frequency bin k. It is known that speech miss detection is more destructive for speech enhancement applications than false alarms. Therefore, we choose the following conservative approach and introduce a second speech absence detector based on $\psi(k,l)$ and the concept of speech presence correlation over neighboring frequency bins that has been exploited in earlier contributions such as [3], [8], [10]. With the help of this second detector, we can decide whether speech is absent based on the local, global, and frame-wise results. We follow the notation of [3] and define the global and frame-based averages of *a posteriori* SNR for the kth frequency bin as

$$\psi_{\mathrm{global}}(k,l) = \sum_{i=-K_1}^{K_1} w_{\mathrm{global}}(i)\psi(k-i,l) \tag{19}$$

where w_{local} is a normalized Hann window of size $2K_1+1$ and

$$\psi_{\mathrm{frame}}(l) = \frac{1}{K}\sum_{i=1}^{K} \psi(i,l). \tag{20}$$

Then, we can decide that the speech is absent in a given frequency bin, i.e., $\hat{q}_{\rm global}(k,l)=1$, if $\psi_{\rm global}(k,l)<\psi_0$, otherwise it is present and $\hat{q}_{\rm global}(k,l)=0$. Similarly, we decide that the speech is absent in the lth frame, i.e., $\hat{q}_{\rm frame}(l)=1$ if $\psi_{\rm frame}(l)<\psi_0$, otherwise it is present and $\hat{q}_{\rm frame}(l)=0$. Finally, we propose the following *a priori* SAP

$$\hat{q}(k,l)=\hat{q}_{\rm local}(k,l)\hat{q}_{\rm global}(k,l)\hat{q}_{\rm frame}(l). \quad (21)$$

Actually, implementation issues may arise when having $\hat{q}(k,l)=1$ as it can be inferred from (9). Therefore, we use $\min\left(\hat{q}(k,l),q_{\max}\right)$ instead of $\hat{q}(k,l)$ when computing the MC-SPP where $q_{\max}=0.99$.

At time frame l, we have an estimate of the noise PSD matrix. Also, we have an estimate of the noisy data PSD matrix that is continuously updated. We use both matrices to obtain an estimatte of the noise-free PSD matrix $\hat{\mathbf{\Phi}}_{xx}(k,l)=\hat{\mathbf{\Phi}}_{yy}(k,l)-\hat{\mathbf{\Phi}}_{vv}(k,l)$. Then, it is straightforward to estimate $\xi(k,l)$ as $\hat{\xi}(k,l)=\mathrm{tr}\left[\hat{\mathbf{\Phi}}_{vv}^{-1}(k,l)\hat{\mathbf{\Phi}}_{xx}(k,l)\right]$. Finally, we implement the proposed MC-SPP estimation approach as a front-end followed by any of the PMWF-based noise reduction methods.

VI. Simulation Results

We consider a simulation setup where a target speech signal composed of six utterances of speech (half male and half female) taken from the IEEE sentences [5], [16] and sampled at 8 kHz rate. The speech signal is convolved with the impulse responses measured off-line at the the Bell-labs acoustic room with a reverberation time $T_{60}=280$ ms. The impulse responses corresponding to different speaker locations and a uniform linear array of 22 microphones in addition to a detailed description of the room configuration are available online in [17]. We assume that the desired speaker is located at "v25" while the interference is located at "v23." We consider the case where the first 2 and 4 microphones only are used for speech acquisition. Two different types of noise are studied: interference (nonspeech taken from the noisex database [18]) from a point source and a computer generated Gaussian noise. The levels of the two types of noise are controlled by the signal-to-interference ratio (SIR) and SNR depending on the scenarios investigated below. To implement the proposed algorithm we choose a frame width of 32 ms for the anechoic environment and 64 ms for the reverberant one in order to capture the long channel impulse response, with an overlap of 50% and a Hamming window for data framing. The filtered signal is finally synthesized using the overlap-add technique. We also choose a Hann window for $w_{\rm global}$, $K_1=15$, $L=32$, $\alpha_p=0.6$, and $\alpha_v=\alpha_y=0.92$.

The results are presented for three types of interfering signals: F-16 and babble, in addition to the case of white Gaussian noise. The SIR is chosen as SIR $=5$ dB. Also a computer generated white Gaussian noise was added such that the input SNR $=20$ dB (the overall input SINR ≈ 4.8 dB). Two and four microphones were respectively used to process the data in both anechoic and reverberant environments. Let $v_{\rm residual}(t)$ and $x_{\rm filtered}(t)$ respectively denote the final residual noise-plus-interference and filtered clean speech signal at the output of one of the above three methods (after filtering, inverse Fourier transform, and synthesis). Then, the performance measures that we consider here are [2], [7]

- Output SINR given by $\frac{E\left\{x_{\rm filtered}^2(t)\right\}}{E\left\{v_{\rm residual}^2(t)\right\}}$,
- Noise (plus interference) reduction factor given by $\frac{E\left\{v_1^2(t)\right\}}{E\left\{v_{\rm residual}^2(t)\right\}}$,
- Signal distortion index given by $\frac{E\left\{[x_1(t)-x_{\rm filtered}(t)]^2\right\}}{E\left\{x_1^2(t)\right\}}$.

For better illustration, we choose three particular values for $\mu=0,1$, and 5 in the PMWF expression.

Notice, first, the important gains in terms of noise reduction when using more microphones in either reverberant or anechoic environments. Indeed, using four microphones leads to better speech detection as shown previously and also more noise reduction as expected [2]. The increase of the parameter μ in the PMWF expression results in more gains in terms of noise reduction and even larger output SINR in all scenarios. However, it also causes more distortions of the desired speech signal. These results lend credence to the study in [2]. Furthermore, we see that in all cases, the least noise reduction factor is achieved in the case of babble noise which is highly non-stationary (as compared to the other two types of interference). This happens because the noise statistics vary at a relatively high rate that they become difficult to track and more noise components are left due to estimation errors of the noise PSD matrix.

VII. Conclusion

In this paper, we proposed a new approach to online multichannel noise tracking and reduction for speech communication applications. This approach can be viewed as a natural generalization of the previous single-channel noise tracking and reduction techniques to the multichannel case. We showed that the principle of MCRA can be extended to the multichannel case. Based on the Gaussian statistical model assumption, we formulated the MC-SPP and combined it with a noise estimator using a temporal smoothing. Then, we developed a two-iteration procedure for accurate detection of speech components and tracking of non-stationary noise. Finally, the estimated noise PSD matrix and MC-SPP were utilized for noise reduction. Good performance in terms of speech detection, noise tracking and speech denoising were obtained.

References

[1] M. Souden, J. Chen, J. Benesty, and S. Affes, "Gaussian model-based multichannel speech presence probability," *IEEE Trans. Audio, Speech, Lang. Process.*, pp. 1–12, in press 2010.

[2] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, pp. 260–276, Feb. 2010.

[3] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Process. Lett.*, vol. 9, pp. 113–116, Apr. 2002.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, pp. 1109–1121, Dec. 1984.

[5] P. C. Loizou, *Speech enhancement: Theory and Practice*. New York, USA: CRC Press, 2007.

	Filter	MVDR			Wiener			Modified Wiener		
	Interf. Sig.	F-16	Babble	Tank	F-16	Babble	Tank	F-16	Babble	Tank
2 Mics.	Output SINR	15.36	14.21	15.69	16.49	14.21	17.14	18.50	16.42	19.40
	Noise reduction factor	10.97	8.60	11.30	12.19	9.89	12.87	14.49	12.37	15.43
	Signal distortion index	−14.72	−15.02	−14.92	−14.70	−14.96	−14.89	−13.96	−14.22	−13.84
4 Mics.	Output SINR	21.14	17.44	18.92	22.15	18.76	20.56	23.68	21.10	22.95
	Noise reduction factor	16.88	13.15	14.65	17.92	14.52	16.35	19.59	17.00	18.93
	Signal distortion index	−14.53	−14.81	−14.84	−14.51	−14.80	−14.85	−14.36	−14.51	−14.38

TABLE I

PERFORMANCE OF THE THREE NOISE REDUCTION FILTERS CORRESPONDING TO $\beta = 0, 1,$ AND 5 IN DIFFERENT NOISE CONDITIONS, INPUT SNR $= 20$ DB, INPUT SIR $= 5$ DB (INPUT SINR≈ 4.8 DB). REVERBERANT ROOM. ALL MEASURES ARE IN DB.

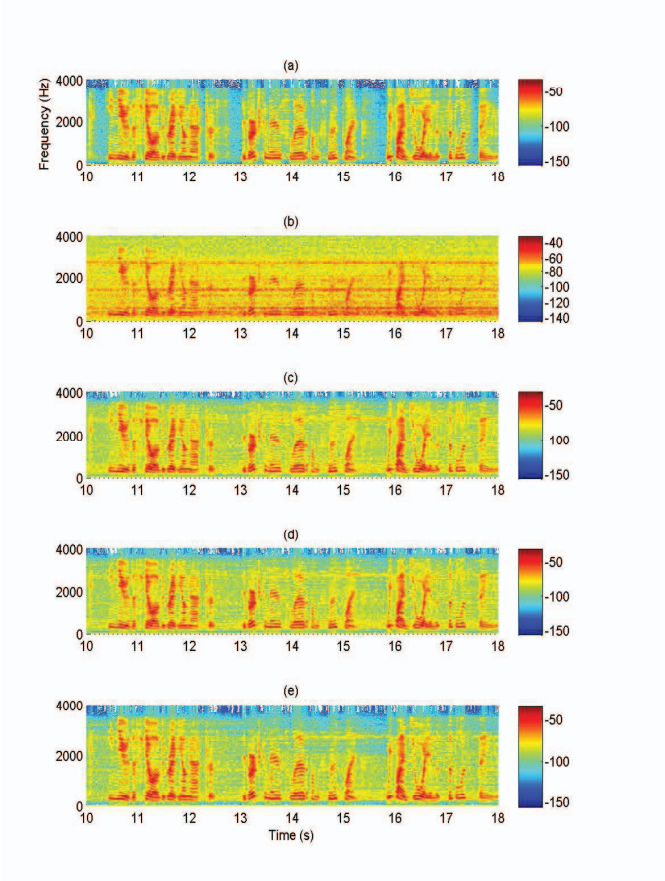

Fig. 1. Spectrograms of portions of the (a) desired clean speech, (b) noisy speech, (c) MVDR (PMWF–$\mu = 0$) output, (d) Wiener (PMWF–$\mu = 1$) filter output, (e) PMWF–$\mu = 5$ output.

[6] I. Y. Soon, S. N. Koh, and C. K. Yeo, "Improved noise suppression filter using self-adaptive estimator for probability of speech absence," *Elsevier, Signal Process.*, vol. 75, pp. 151–159, Sep. 1999.

[7] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer-Verlag, 2008.

[8] S. Gannot and I. Cohen, *Springer Handbook of Speech Processing*, ch. Adaptive beamforming and postfitering, pp. 945–978. Berlin, Germany: Springer-Verlag, 2007.

[9] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, pp. 504–512, Jul. 2001.

[10] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, pp. 466–475, Sept. 2003.

[11] D. Middleton and R. Esposito, "Simultaneous optimum detection and estimation of signals in noise," *IEEE Trans. Inf. Theory*, vol. IT-14, pp. 434–444, May 1968.

[12] I. Potamitis, "Estimation of speech presence probability in the field of microphone array," *IEEE Signal Process. Lett.*, vol. 11, pp. 956–959, Dec. 2004.

[13] G. A. F. Seber, *Multivariate Distributions*. New York, USA: John Wiley & Sons, Inc., 1984.

[14] I. S. Gradshteyn and I. Ryzhik, *Table of Integrals, Series, and Products, Seventh Edition*. Elsevier Acad. Press, 2007.

[15] J. J. McKeon, "F approximations to the distribution of Hotelling's T_0^2," *Biometrika*, vol. 61, pp. 381–383, Aug. 1974.

[16] IEEE, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoustics*, vol. 17, pp. 225–246, 1969.

[17] "http://www.acoustics.hut.fi/ aqi/vardata/varechoic_array_data.html."

[18] A. P. Varga, H. J. M. Steenekan, M. Tomlinson, and D. Jones, "The noisex-92 study on the effect of additive noise on automatic speech recognition," tech. rep., DRA Speech Research Unit, 1992.