A New Algorithm for Noise PSD Matrix Estimation in Multi-Microphone Speech Enhancement Based on Recursive Smoothing

Mahdi Parchami*, Wei-Ping Zhu* and Benoit Champagne†

*Department of Electrical and Computer Engineering
Concordia University, Montreal, Quebec, Canada H3G 1M8
Emails: m_parch@ece.concordia.ca, weiping@ece.concordia.ca

†Department of Electrical and Computer Engineering
McGill University, Montreal, Quebec, Canada H3A 0E9
Email: benoit.champagne@mcgill.ca

***Abstract*—In this paper, we present a new algorithm for the estimation of the noise power spectral density (PSD) matrix, as needed for multi-microphone speech enhancement in a general non-stationary noisy environment. First, we propose a recursive scheme for noise PSD estimation in which the current, previous and close subsequent noisy speech frames are properly weighted. The forgetting factor for the recursive updating of the smoothed PSD is obtained based on an overall measure of the SNR across all microphone signals. Since this SNR measure depends on the noise statistics, we choose to iteratively update it using the latest available estimate of the noise PSD matrix. Finally, to obtain better estimation accuracy in the proposed method, we further apply a direct extension of the minimum tracking approach to the estimated noise PSD matrix. Performance of the proposed algorithm is evaluated in terms of objective measures and its superiority is shown with respect to two recent noise PSD estimation methods in the context of speech enhancement.**

***Keywords*—Microphone array, noise PSD matrix estimation, noise reduction, speech enhancement.**

I. Introduction

Single-channel frequency-domain speech enhancement systems aim at suppressing the background noise by modifying the spectrum of noisy speech signal without degrading the quality of the desired speech. In this regard, the estimation of noise power spectral density (PSD) is often required and its accuracy highly influences the quality of the enhanced speech [1]. To achieve further noise reduction while keeping the speech distortion at a minimum level, multi-channel speech enhancement methods have been proposed that apply optimal beamforming to the observations of a microphone array. These methods, however, in order to fully exploit the spatial diversity among different microphone observations require the *a priori* knowledge of the spatial noise PSD matrix which is generally challenging to estimate [2].

Recently, considerable research has been directed toward the estimation of the noise PSD matrix. In this regard, due to the popularity of the groundbreaking method of minimum statistics (MS) proposed by Martin [3], a few straightforward multi-channel extensions of this method have been presented. In [4], a two-channel noise PSD estimator has been suggested by combining the MS method and a voice activity detector (VAD). However, the VAD-based noise estimation techniques are not capable of providing as much accuracy as the soft-decision methods, due to the lack of noise PSD updating during frames where the speech component is present [5]. In [6], an MS-based method to estimate the noise PSD matrix has been proposed by using the recursive smoothing of noisy speech through a fixed forgetting factor. However, as proved in the context of single-channel noise estimation, selecting the forgetting factor independently for each frame/frequency can largely enhance the noise estimation accuracy. In [7], an algorithm for the estimation of the noise PSD matrix has been suggested by employing an adaptive forgetting factor selected based on multi-channel speech presence probability (SPP). Yet, the SPP employed in [7] is obtained under a two-hypotheses basis assuming either the presence or the absence of speech in all channels, which is not accurately true due to the difference among the observations in different channels. Another recent method has been proposed in [8] where it is attempted to eliminate the undesirable speech component while estimating the noise PSD matrix. Nevertheless, due to employing the conventional fixed smoothing in its structure, it results in trivial improvements at moderate SNRs.

In this work, we first make use of subsequent speech frames to achieve a more efficient smoothing scheme on noisy observations. We then update the forgetting factor in an iterative manner by taking into account the overall SNR in microphone channels. Lastly, minimum tracking is further applied to the estimated noise PSD matrix. We evaluate the performance of the proposed algorithm in comparison with other major methods in the context of speech enhancement.

II. Problem Formulation

Let $s(t)$ denote a source of speech signal in a noisy environment impinging on an array of N microphones with an arbitrary geometry at time instant t. The resulting observation at the nth microphone can be written as

$$y_n(t) = x_n(t) + v_n(t), \quad n = 1, 2, ..., N \tag{1}$$

978-1-4799-8391-9/15/$31.00 ©2015 IEEE

with $x_n(t)$ and $v_n(t)$ being the received speech and the noise at the nth microphone. The speech term, $x_n(t)$, can be considered as $s(t) * g_n(t)$ with $g_n(t)$ denoting the channel impulse response from the speech source to the nth microphone and $*$ as the convolution operator. It is assumed that the noise terms, $v_n(t)$, are uncorrelated with $x_n(t)$. Therefore, in the short-time Fourier transform (STFT) domain, the signal model in (1) is

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

with k and l denoting respectively the frequency bin and time frame indices and $G_n(k,l)$ as the nth channel frequency response. By expressing (2) in the vector form, we have

$$\mathbf{Y}(k,l) = S(k,l)\mathbf{G}(k,l) + \mathbf{V}(k,l) \quad (3)$$

The aim of spectral domain speech enhancement is to provide an estimate of the clean speech spectrum, $S(k,l)$, given the set of noisy observations, $\mathbf{Y}(k,l)$. Given the above, the noise PSD matrix is defined as $\Sigma_{\mathbf{VV}} \triangleq E\{\mathbf{V}\mathbf{V}^H\}$ with $E\{.\}$ denoting the expectation and H the hermitian transpose. The aim of this work is to provide an estimate for $\Sigma_{\mathbf{VV}}$ that is to be used in microphone array beamforming techniques.

III. Proposed Algorithm for Noise PSD Matrix Estimation

In this section, we propose a new noise PSD matrix estimation consisting of a new recursive smoothing scheme and an extension to the minimum tracking.

A. Incorporation of Subsequent Speech Frames

All prior solutions to noise estimation problem include recursive smoothing schemes using the current and past noisy speech frames. This is due to the need for ensemble averaging implied by the statistical expectation $E\{.\}$. In this sense, to make use of all the available information, we suggest to take advantage of several following speech frames in the recursive smoothing performed for the noise PSD estimation. On this basis, we propose the following weighted recursive smoothing scheme for the estimation of noise PSD matrix

$$\begin{aligned}\mathbf{P}(k,l) = \beta\ \alpha(k,l)\mathbf{P}(k,l-1) + [1-\alpha(k,l)]\,\mathbf{Y}(k,l)\mathbf{Y}^H(k,l) \\ +(1-\beta)\ \alpha(k,l)\sum_{i=1}^{D} w_i \mathbf{Y}(k,l+i)\mathbf{Y}^H(k,l+i)\end{aligned} \quad (4)$$

where $\mathbf{P}(k,l)$ is the smoothed noisy spectrum, $\alpha(k,l)$ is the forgetting factor in smoothing the past frames, β is the smoothing parameter used to determine the weighting between the past and future frames and w_i are the weighting scheme applied on the D future frames. It should be noted that the exploitation of D future frames in the noise estimation for current speech frame implies a certain processing delay. Yet, due to the practical range of D, say $D \leq 5$, and the overlap between consecutive frames, the amount of delay is negligible as it is smaller than a few decades of milliseconds only. As for the weighting parameter β, an experimentally fixed value of 0.65 has worked best in the tested scenarios, which gives more emphasis to the numerous past frames. The selection of $\alpha(k,l)$ will be discussed in the following subsection. As for the weightings w_i, we consider a fixed exponential scheme as $w_i = \gamma^i$, noting that the conventional recursive smoothing performed on past frames results in an exponential scheme for its weightings (as eq. (13) in [3]). Given this and the fact that $\sum_{i=1}^{D} w_i = 1$, we end up with the following equation in terms of γ exponent

$$\gamma^{D+1} - 2\gamma + 1 = 0, \quad \text{for known } D \quad (5)$$

It should be noted that for small D values, (5) has exactly one real-valued positive solution that makes it possible to use γ^i as a proper weighting.

B. Iterative Method for the Selection of Forgetting Factor

In spite of the high importance in the selection of the forgetting factor, $\alpha(k,l)$, the literature on noise PSD matrix estimation lacks efficient schemes for this purpose. We herein take into account the fact that, in the recursive smoothing of noisy speech, a larger weight should be assigned to the update term when the speech component is weaker (or equivalently the noise component is stronger) and vice versa [5]. To this end, we suggest to measure the speech signal intensity in all channels by the following definition of the overall SNR

$$\zeta(k,l) \triangleq \frac{\left\|\Sigma_{\mathbf{XX}}(k,l)\right\|_2}{\left\|\Sigma_{\mathbf{VV}}(k,l)\right\|_2} = \frac{\left\|\Sigma_{\mathbf{YY}}(k,l) - \Sigma_{\mathbf{VV}}(k,l)\right\|_2}{\left\|\Sigma_{\mathbf{VV}}(k,l)\right\|_2} \quad (6)$$

where $\Sigma_{\mathbf{XX}}$ denotes the speech PSD matrix, with $\mathbf{X}(k,l)$ defined as $S(k,l)\mathbf{G}(k,l)$, and the notation $\|.\|_2$ indicates the L_2-norm of a matrix. The equation at the right of (6) holds due to the uncorrelated speech and noise components. Based on this measure of SNR, we propose to select the forgetting factor as

$$\alpha(k,l) = \alpha_{min} + (\alpha_{max} - \alpha_{min})\,\zeta_{norm}(k,l) \quad (7)$$

with α_{min} and α_{max} as the fixed minimum and maximum values for $\alpha(k,l)$ chosen as 0.25 and 0.94, respectively, and $\zeta_{norm}(k,l)$ is the thresholded and normalized $\zeta(k,l)$ given by

$$\zeta_{norm}(k,l) = \begin{cases} 1, & \text{if } \zeta(k,l) \geq T_H \\ \frac{\zeta(k,l) - T_L}{T_H - T_L}, & \text{if } T_L < \zeta(k,l) < T_H \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

with the high and low thresholds $T_H = 22$ and $T_L = 0.35$, in respect. Now to implement (6), proper estimates of $\Sigma_{\mathbf{YY}}(k,l)$ and $\Sigma_{\mathbf{VV}}(k,l)$ are needed. The PSD matrix of noisy speech, $\Sigma_{\mathbf{YY}}(k,l)$, can be easily estimated through the recursive smoothing of the noisy observations. However, an estimate of $\Sigma_{\mathbf{VV}}(k,l)$ is not available. Therefore, we propose the following iterative algorithm to estimate $\alpha(k,l)$:

(1) Initialization: Use $\mathbf{P}(k,l-1)$ in place of $\Sigma_{\mathbf{VV}}(k,l)$.
(2) Calculate $\zeta(k,l)$ using (6).
(3) Use $\zeta(k,l)$ to obtain $\zeta_{norm}(k,l)$ in (8).
(4) Calculate $\alpha(k,l)$ using (7).
(5) Use $\alpha(k,l)$ to obtain $\mathbf{P}(k,l)$ in (4)
(6) Use $\mathbf{P}(k,l)$ in place of $\Sigma_{\mathbf{VV}}(k,l)$ in step *(2)*.
(7) Continue the next steps.

with $\mathbf{P}(k,l)$ taken as the estimate for the noise PSD matrix at the end of each iteration. As for the first frame, assuming that there is no speech component present, $\alpha(k,1)$ is chosen as α_{min} and then $\mathbf{P}(k,1)$ is calculated. In all the investigated noise scenarios, we found that using only two iterations of the

above was sufficient and no considerable improvements were obtained if using more iterations.

C. Minimum Tracking and Bias Compensation

We here employ an extension of the minimum tracking method [3] to further improve the accuracy of noise PSD estimation. To this end, we track the minimum norm of the noise PSD matrix estimate, i.e. $\mathbf{P}(k,l)$, across the current and last $M-1$ frames. Therefore, we define $\mathbf{P}_{min}(k,l)$ as the matrix with minimum L_2-norm on the set $\{\mathbf{P}(k,l), \mathbf{P}(k,l-1), ..., \mathbf{P}(k,l-M+1)\}$. Yet, as stated in [3], $\mathbf{P}_{min}(k,l)$ is biased toward lower values and the bias needs to be compensated. Based on the statistics of the minimum tracking, this bias has been estimated in [3] for the case of noise PSD estimation. However, the problem becomes theoretically too tedious when dealing with noise PSD matrix estimation. For this reason, considering that the bias is linearly dependent on the number of frames, M, as evident in eq. (17) in [3], we found the following approximation to the inherent bias in $\mathbf{P}_{min}(k,l)$ to be useful

$$B_{min} \approx 1 + \frac{M-1}{2} \tag{9}$$

Now by devision of the minimum tracked value, $\mathbf{P}_{min}(k,l)$, by its bias in the above, we obtain the ultimate estimate for the noise PSD matrix as

$$\hat{\Sigma}_{\mathbf{VV}}(k,l) = \frac{\mathbf{P}_{min}(k,l)}{B_{min}} \tag{10}$$

The above is to be used as the proposed estimate for the noise PSD matrix in the following section.

IV. Performance Evaluation

In this section, we investigate the performance of the proposed algorithm for the estimation of noise PSD matrix in non-stationary noise scenarios. Clean speech sentences from both male and female speakers were chosen from TIMIT database [9] and different noise types were added from NOISEX-92 [10]. The sampling rate was set to 16 kHz and a time segmental length of 20 ms with 75% of overlapping between consecutive Hamming windows were considered for the STFT. To synthesize microphone array signals, Image Source method (ISM) was used [11] with a room dimension of 3m×4m×2m and an $N = 2$ microphone array with an inter-microphone distance of 6 cm. For each of the noise scenarios, i.e. white and babble, two point-source noises were considered impinging on the microphone array from directions of 30 and 120 degrees with respect to the array line. Also, the single speech source was assumed to be located on the microphone array boresight. In all simulations, the number of subsequent frames considered in the smoothing was assumed to be $D = 3$ implying that $\gamma =$ 0.5437. Even though small improvements were obtainable by increasing D up to 5, for the sake of comparable complexity burden, we kept D at 3. This also ensures the imposed processing delay not to be more than 15 ms. We evaluate the proposed noise PSD matrix estimation algorithm in comparison with two recent methods in the speech enhancement literature, namely Souden's method in [7] and Hendriks' method in [8], and also the noise PSD matrix obtained by smoothing noise only samples. The latter can be counted on as a close approximation

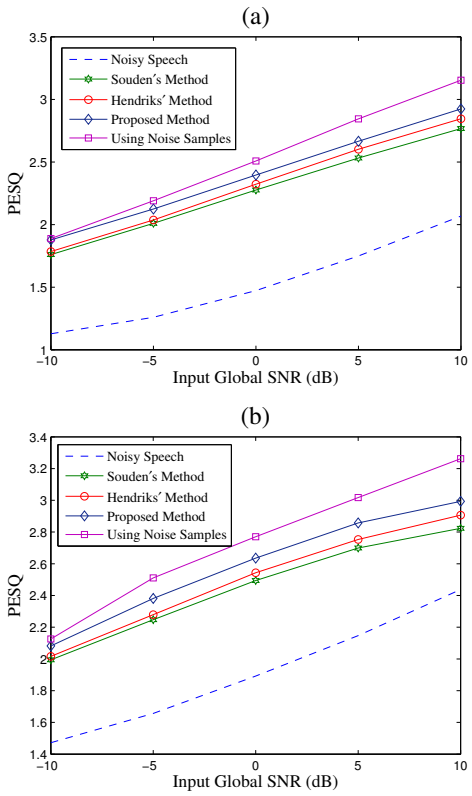

Fig. 1. PESQ scores versus input global SNR using MVDR beamformer with different noise PSD matrix estimates for: (a) white noise, (b) babble noise.

to an ideal estimate for the noise PSD matrix, serving as an upper bound for the estimation methods.

First, we implemented the minimum variance distortionless response (MVDR) beamformer using the noise PSD matrix estimated by the aforementioned methods and assessed the performance in terms of two objective measures, PESQ and segmental SNR. Whereas the former measures the total speech quality, the latter gives a perspective of the amount of noise reduction. Fig. 1 depicts PESQ scores using different noise PSD matrix estimation methods for the input global SNR in the range of -10 dB to 10 dB. It is observable that the proposed algorithm outperforms the other two methods in almost all of the input SNR range by over 0.1, which is a considerable improvement in the speech quality. Note that there is still a large gap between the employed methods and that using the noise only samples to estimate the noise PSD matrix, especially in higher input SNRs. The reason is due to the presence of the strong speech components in the estimated elements across the noise PSD matrix in the soft-decision based methods, which results in speech signal cancellation and unfavorable distortion in the MVDR output. Fig. 2 shows the same trend using the segmental SNR measure. It is visible that improvements of almost 1∼2 dBs are obtained by employing the proposed method relative to the previous approaches. To further evaluate the performance of the noise PSD matrix estimation methods, we plotted the MVDR beamformer response (output) errors in Fig. 3, as suggested by equation (37) in [8]. This in fact shows a measure of distance between the reference output obtained through the noise only samples and the outputs by the other methods. Due to the smaller beamformer response error in the

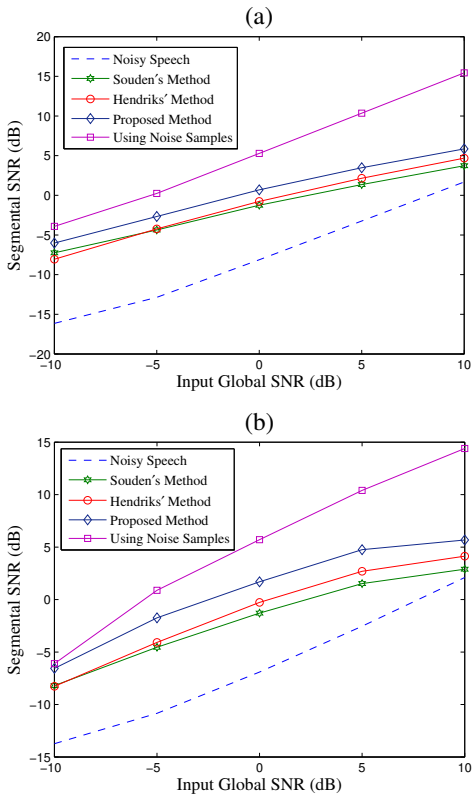

Fig. 2. Segmental SNR of the enhanced speech versus input global SNR using MVDR beamformer with different noise PSD matrix estimates for: (a) white noise, (b) babble noise.

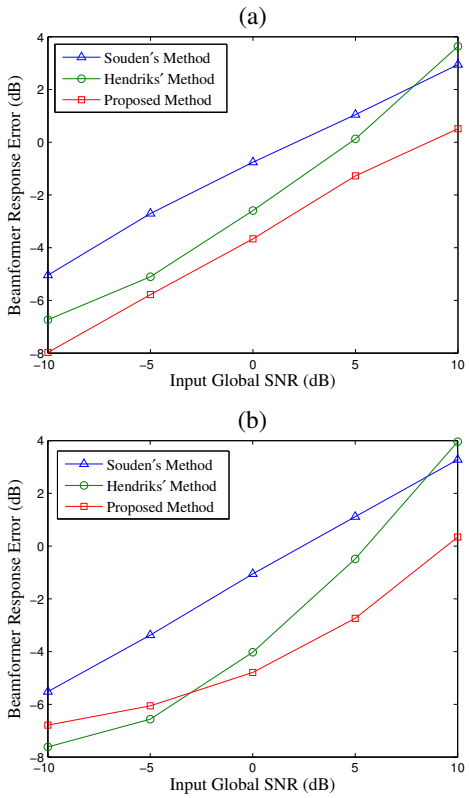

Fig. 3. MVDR beamformer response error versus input global SNR using different noise PSD matrix estimates for: (a) white noise, (b) babble noise.

proposed method, as observed in Fig. 3, it is apparent that the proposed algorithm attains an MVDR output closer to that of the reference method. Evaluations with respect to other types of non-stationary noise were also performed, confirming the superiority of the proposed algorithm in all scenarios.

V. Conclusions

We proposed an algorithm for the estimation of noise PSD matrix in a non-stationary noisy field. The presented algorithm employs subsequent noisy speech frames, in addition to previous frames, to update the noise PSD estimate for the current frame. The forgetting factor employed in the smoothing scheme is updated iteratively as a function of the past noise PSD matrix estimate and is therefore adaptively adjusted to the last available estimate of the noise PSD matrix. A minimum-norm tracking scheme is also suggested for the estimated noise PSD matrix to further enhance the estimation accuracy. Performance evaluations were performed by using the noise PSD matrix estimates obtained from the proposed and two recent approaches in the MVDR beamformer and the advantage of the proposed algorithm over the past two methods was confirmed.

ACKNOWLEDGMENT

This work was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada, and the Regroupement Stratgique en Microlectronique du Qubec (ReSMiQ).

References

[1] P.C. Loizou, "*Speech Enhancement: Theory and Practice*," CRC Press, 2007.

[2] J. Benesty, J. Chen and Y. Huang, "*Microphone Array Signal Processing*," Springer Science & Business Media, 2008.

[3] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. on Speech and Audio Process.*, Vol. 9, No. 5, pp. 504-512, July 2001.

[4] J. Freudenberger, S. Stenzel and B. Venditti, "A noise PSD and cross-PSD estimation for two-microphone speech enhancement systems," in *15th Workshop on Statistical Signal Processing*, Aug. 2009.

[5] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Trans. on Speech and Audio Process.*, Vol. 11, No. 5, pp. 466-475, Sep. 2003.

[6] F. Kallel et al., "A noise cross-PSD estimator based on improved minimum statistics method for two-microphone speech enhancement dedicated to a bilateral cochlear implant," *Applied Acoustics*, Vol. 73, No. 3, pp. 256-264, 2012.

[7] M. Souden, J. Chen, J. Benesty, S. Affes, "An integrated solution for online multichannel noise tracking and reduction," *IEEE Trans. on Audio, Speech and Language Process.*, Vol. 19, No. 7, pp. 2159-2169, Sep. 2011.

[8] R.C. Hendriks, T. Gerkmann, "Noise correlation matrix estimation for multi-Microphone speech enhancement," *IEEE Trans. on Audio, Speech and Language Process.*, Vol. 20, No. 1, pp. 223-233, Jan. 2012.

[9] J.S. Garofolo, "DARPA TIMIT acoustic-phonetic speech database," National Institute of Standards and Technology (NIST), 1988.

[10] Noisex-92 database, *Speech at CMU*, Carnegie Mellon University, available at: http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html.

[11] E. Lehmann and A. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *Journal of the Acoustical Society of America*, Vol. 124(1), pp. 269-277, July 2008.