NOISE POWER SPECTRAL DENSITY MATRIX ESTIMATION BASED ON MODIFIED IMCRA

Qipeng Gong, Benoit Champagne and Peter Kabal

Department of Electrical & Computer Engineering, McGill University
3480 University St., Montreal, Quebec, Canada H3A 0E9
qi.gong@mail.mcgill.ca, benoit.champagne@mcgill.ca, peter.kabal@mcgill.ca

ABSTRACT

In this paper, we present a new method for noise power spectral density (PSD) matrix estimation based on IMCRA which consists of two parts. For the auto-PSD (diagonal) estimation, we propose a modification to IMCRA where a special level detector is employed to improve the tracking of non-stationary noise backgrounds. For the cross-PSD (off-diagonal) estimation, we propose to calculate a smoothed cross-periodogram by using estimated noise components derived as residuals after the application of a speech enhancement algorithm on the individual microphone signals. Simulation results show the effectiveness of our proposed approach in estimating the noise PSD matrix and its robustness against reverberation when used in combination with an MVDR-based speech enhancement system.

1. INTRODUCTION

In voice communication systems, the speech signal on the transmitter side is often corrupted by various types of background acoustic noise. To obtain a high quality speech signal on the receiver side, it is desired to reduce the noise level without introducing noticeable distortion to the target speech, or worst, affecting its intelligibility. To this end, since we do not have access to the background noise signal, it is necessary to use information about the statistical characteristics of the noise, especially its second order moments in the form of the noise power spectral density (PSD).

Existing speech enhancement approaches can be divided into two main classes depending on whether they employ a single microphone (SM) versus a microphone array (MA). In SM approaches, the noise PSD is typically employed to calculate a spectral gain, which in turn is applied to the noisy speech in the frequency domain to obtain the enhanced speech [1]. Traditionally, noise PSD estimation has been based on voice activity detectors (VADs), which restrict the update of the PSD estimate to periods of speech absence. However, VADs are often difficult to tune and their reliability deteriorates severely at low signal-to-noise ratio (SNR). In recent years, alternative estimation approaches have therefore been proposed that do not directly rely on VAD. In [2], a noise PSD estimator based on minimum statistics (MS) is studied, which tracts the minima values of a smoothed PSD estimate of the noisy signal and multiplies the result by a bias factor. In the so-called improved minima controlled recursive averaging (IMCRA) [3], smoothing of the noisy speech periodogram is controlled by the conditional speech presence probability, which in turn is estimated based on the results of minimum tracking iterations. The advantages of IMCRA are particularly notable in adverse environments involving non-stationary noise and low input SNR.

The use of MA offers many appealing advantages over SM in speech enhancement, including the possibility of realizing distortionless noise reduction through additional degrees of freedom and added flexibility in handling different types of interference, such as multiple talker and reverberation [4]. As in the SM case, the performance of MA techniques strongly depends on side information, especially *a priori* knowledge of the PSD matrix of the background noise and interference. For instance, the PSD matrix plays a key role in the realization of the miminum variance distortionless response (MVDR) beamformer and the multi-channel Wiener filter. However, estimation of the noise PSD matrix, which consists of auto-PSD (diagonal) and cross-PSD (off-diagonal) elements, is much more challenging than that of its SM counterpart. The current literature on PSD matrix estimation for acoustic noise is scarce. In [5, 6], an energy-based VAD is used to enable the cross-PDS estimation only during speech pauses. Other recent methods exploit additional assumptions on the acoustic field, such as diffuse spherically isotropic noise [7] or known propagation vector of the clean speech [8]. However, these assumptions are not always realistic and thus impose severe practical limitations.

In this paper, we present and investigate an improved method for noise PSD matrix estimation based on IMCRA which consists of two parts. For the auto-PSD estimation, we propose a modification to IMCRA where a frequency dependent level detector is employed to improve the tracking of non-stationary noise backgrounds. For the cross-PSD estimation, we propose to calculate the smoothed cross-periodogram by using estimated noise components, derived

[1]Funding for this work was provided by a CRD grant from NSERC (Govt. of Canada) under the sponsorship of Microsemi Corporation (Ottawa, Ontario, Canada).

as residuals following the application of a selected single channel speech enhancement algorithm on the individual microphone signals. Simulation results show the effectiveness of our proposed approach in estimating the noise PSD matrix, and its robustness against reverberation when used in a speech enhancement system based on MVDR beamforming.

This paper is organized as follows: Section 2 presents the notations and problem formulation. The auto-PSD estimator is discussed in Section 3, where we first review IMCRA and then propose a modification to improve its tracking ability. The new IMCRA-based cross-PSD estimator, which employs estimates of the noise components in the microphone signals, is presented in Section 4. Simulation results are presented in Section 5, which is followed by a conclusion in Section 6.

2. PROBLEM FORMULATION

Let us consider an array of M microphones deployed in a noisy environment in which the noise and desired speech signals are spatially separated. The noisy speech signal samples received at the μ-th microphone, $\mu \in \{1, ..., M\}$, can be expressed as

$$y_\mu[m] = s_\mu[m] + n_\mu[m] \tag{1}$$

where $s_\mu[m]$ is the speech component, $n_\mu[m]$ is the additive noise and m is the discrete-time index. Standard short-time Fourier transform (STFT) analysis is applied to the microphone signals, which are synchronously segmented into overlapping frames of length L and frame advance R. The signal samples in each frame are multiplied by an analysis window, denoted as $w(l)$, and then mapped to the frequency domain via the discrete Fourier transform, that is:

$$Y_\mu(k,i) = \sum_{l=0}^{L-1} y_\mu(iR+l)w(l)e^{-j2\pi kl/L} \tag{2}$$

where $Y_\mu(k,i)$ denotes the STFT coefficient of the noisy speech for frequency bin k, time-frame i and microphone μ. Accordingly, in the time-frequency domain, (1) can be expressed as

$$Y_\mu(k,i) = S_\mu(k,i) + N_\mu(k,i) \tag{3}$$

where $S_\mu(k,i)$ and $N_\mu(k,i)$ denote the corresponding STFT coefficients of the speech and noise, respectively.

We model $S_\mu(k,i)$ and $N_\mu(k,i)$ as zero-mean complex random variables, uncorrelated across time and frequency; we also assume that the signal and noise components are mutually independent. In this work, our main interest lies in the second order statistical properties of the noise STFT, as represented by the short-time PSD. Specifically, for the time-frequency point (k,i), let us define

$$P_{\mu,\nu}(k,i) = E\{N_\mu(k,i)N_\nu^*(k,i)\} \tag{4}$$

where $E\{\cdot\}$ denotes expectation and superscript * indicates complex conjugation. In the case $\mu = \nu$, $P_{\mu,\nu}(k,i)$ in (4) is known as the auto-PSD, while if $\mu \neq \nu$, it is called cross-PSD. Accordingly, the noise PSD matrix can be defined as

$$\mathbf{P}(k,i) = \begin{bmatrix} P_{1,1}(k,i) & \cdots & P_{1,M}(k,i) \\ \vdots & \ddots & \vdots \\ P_{M,1}(k,i) & \cdots & P_{M,M}(k,i) \end{bmatrix}. \tag{5}$$

The PSD matrix (5) plays a key role in MA-based speech enhancement. For some algorithms, such as the MVDR beamformer and the multi-channel Wiener filter, this matrix directly determines the spatial filtering being applied to the microphone signals. For instance, the information contained in $\mathbf{P}(k,i)$ makes it possible to steer a MVDR beamformer in the direction of a desired speaker while canceling, or reducing the effect of noise from other directions. Similar to the noise PSD in SM approaches, $\mathbf{P}(k,i)$ needs to be estimated from the noisy microphone signals, and the accuracy of this estimation may greatly affect the performance of the enhancement algorithm. In particular, poor estimation can lead to a situation where disturbances from certain directions are not optimally suppressed, or worse, are amplified by MA processing [8]. Estimation of the noise PSD matrix is challenging, not only because of the speech presence and the noise non-stationarity as in the SM case, but also because of the additional complexity induced by the spatial dimension.

According to (5), we note that the diagonal elements of the noise PSD matrix, i.e., $P_{\mu,\mu}(k,i)$, are ordinary auto-PSD and therefore, methods developed for SM are often applied for their estimation in MA systems. Regarding the off-diagonal elements or cross-PSD, i.e. $P_{\mu,\nu}(k,i)$ for $\mu \neq \nu$, their estimation can also be approached via recursive averaging, as in [5, 6]. Below, we propose improved methods based on IMCRA for the estimation of both the diagonal and off-diagonal elements of the noise PSD matrix.

3. AUTO-PSD ESTIMATOR

3.1. Overview of IMCRA

In IMCRA [3], the noise PSD estimate is obtained by recursively averaging past spectral power values of the noisy speech, using a smoothing parameter which is adjusted by the speech presence probability in each frequency bin. Mathematically, this process for estimating the auto-PSD for the μ-th microphone can be expressed as

$$\hat{P}_{\mu,\mu}(k,i) = \tilde{\alpha}_\mu(k,i)\hat{P}_{\mu,\mu}(k,i-1) + (1-\tilde{\alpha}_\mu(k,i))|Y_\mu(k,i)|^2 \tag{6}$$

where

$$\tilde{\alpha}_\mu(k,i) = \alpha + (1-\alpha)p_\mu(k,i) \tag{7}$$

is the time-varying frequency-dependent smoothing parameter, $p_\mu(k,i)$ is the speech presence probability conditioned on $|Y_\mu(k,i)|^2$ and α is a (fixed) secondary smoothing parameter.

In a conventional VAD-based algorithm, the noise PSD would be estimated recursively with smoothing parameter α when speech is absent, and held constant when it is present. In contrast, the auto-PSD estimation by IMCRA depends on a "soft" decision, namely the conditional speech presence probability $p_\mu(k,i)$, instead of a binary VAD indicator. In effect, the noise PSD is continually adapted based on the noisy measurements and the smoothing parameter $\tilde{\alpha}_\mu(k,i)$ is changed accordingly, i.e. being increased when $p_\mu(k,i)$ is large and vice versa. This makes it possible to adjust the integration time of the estimator depending on the speech activity in each frequency bin over time.

The speech presence probability is generally biased toward higher values to avoid speech distortion in speech enhancement applications. Consequently, the auto-PSD estimation based on recursive averaging would be biased toward lower values. To offset this effect, a multiplicative bias compensation factor $\beta > 1$ is usually applied to the PSD estimator (6), whose value can be determined based on theoretical considerations but is often set to around 1.5 in practice.

The expression of the conditional speech presence probability $p_\mu(k,i)$ in (7) can be obtained based on a Gaussian statistical model. Specifically, let us define the *a posteriori* and *a priori* SNR as follows, respectively:

$$\gamma_\mu(k,i) = \frac{|Y_\mu(k,i)|^2}{P_{\mu,\mu}(k,i)}, \quad \xi_\mu(k,i) = \frac{E\{|S_\mu(k,i)|^2\}}{P_{\mu,\mu}(k,i)}. \quad (8)$$

In terms of these quantities, we have

$$p_\mu(k,i) = \Big(1 + \frac{q_\mu(k,i)(1+\xi_\mu(k,i))}{1-q_\mu(k,i)} e^{-\frac{\gamma_\mu(k,i)\xi_\mu(k,i)}{1+\xi_\mu(k,i)}}\Big)^{-1} \quad (9)$$

where $q_\mu(k,i)$ is the *a priori* probability for speech absence, which is controlled by the result of the minimum tracking. Specifically, two iterations of smoothing and minimum tracking are employed in IMCRA to estimate $q_\mu(k,i)$: The first one provides a rough VAD in each frequency bin while the second one excludes relatively strong speech components, for added robustness in the minimum tracking during speech activity. The details of this process can be found in [3].

3.2. Proposed Modification to IMCRA

When using IMCRA, a large estimation error may occur after an abrupt increase in the noise level. In the past, some improvements have been suggested to reduce this tracking delay, e.g. [9]. Here, we present a simple yet effective scheme based on energy detection which exploits the different spectral distributions of the speech and noise power.

The slow response time of IMCRA stems from the strategy used to update the search window for the minimum tracking, which must employ a somewhat too long memory of past input frames. In theory, the problem can be resolved by firstly detecting the level increment in the background noise power and then resetting the search window with data from the current frame. To this end, we propose a noise increment detector based on monitoring changes in both the high and low frequency power content of the noisy speech, which is motivated as follows. When speech is present, a detected power level increment in the noisy speech could be the result of a sudden increase in the power level of the desired speech. Still, we notice that the power of a speech signal is mainly localized in a band of frequencies from say 300Hz to 6kHz, while the noise power tend to spread through all the frequency bins. Hence, the changes in the power of the observed noisy speech at lower frequencies (say $f \le f_L = 300$Hz) and higher frequencies ($f > f_H = 6$kHz) are most likely caused by an increase in the background noise level, which can be exploited to avoid false detection. On this basis, we propose to modify IMCRA as follows.

For the μ-th microphone, let us define the instantaneous power of the observed noisy speech within the low and high frequency bands at the i-th frame as follows, respectively:

$$P_\mu^L(i) = \sum_{k=0}^{k_L} |Y_\mu(k,i)|^2, \; P_\mu^H(i) = \sum_{k=k_H}^{L/2-1} |Y_\mu(k,i)|^2 \quad (10)$$

where $k_L = \lfloor \frac{300L}{F_s} \rfloor$, $k_H = \lceil \frac{6000L}{F_s} \rceil$ and F_s is the sampling frequency in Hz. Also define the corresponding increments in power levels over consecutive frames, i.e.: $\Delta P_\mu^L(i) = P_\mu^L(i) - P_\mu^L(i-1)$ and $\Delta P_\mu^H(i) = P_\mu^H(i) - P_\mu^H(i-1)$. The proposed algorithm uses the above differential power measures in combination with two thresholds, denoted by γ_L and γ_H, to detect a sudden increment in the noise level. Specfically, a binary indicator variable is first calculated as follows:

$$\text{Ind}(i) = \begin{cases} 1, & \Delta P_\mu^L(i) > \gamma_H \text{ and } \Delta P_\mu^H(i) > \gamma_L \\ 0, & \text{otherwise} \end{cases} \quad (11)$$

A change from 0 to 1 in Ind(i) indicates a possible sudden increase in the background noise level. However, especially at higher SNR, such a change might be the result of a sudden increase in the power level of the desired speech. To avoid this behavior, i.e. false alarm in the detection of a noise level increment, it is preferable to introduce a timing delay before making a final decision. Specifically, following a change from 0 to 1 in Ind(i), we require that $P_\mu^H(i)$ remains large for a sufficient number of frames, say $n_{\text{fr}} = 6$, before deciding for an increase in the noise level; otherwise the process is stopped. This second test involves a third threshold, which we denote as γ_{stop}.

Finally, following the detection of a sudden increase in the noise level, the IMCRA variables related to minimum tracking are reset to their initial values (i.e., as used for the first frame) in all the frequency bins. The complete procedure is summarized in pseudo-code form in Algorithm 1. In the rest of this paper, we refer to the auto-PSD estimation algorithm that results from incorporating this modification into IMCRA as the *modified* IMCRA.

Algorithm 1 Noise Level Increment Detection

```
Initialize Low_old and High_old
Initialize Ind = 0
for i = 0, 1, ... do
  ΔP_L = P_μ^L(i) − Low_old
  ΔP_H = P_μ^H(i) − High_old
  if Ind == 0 then
    if ΔP_H ≥ γ_H and ΔP_L ≥ γ_L then
      Ind = 1
    else
      High_old = P_μ^H(i)
      Low_old = P_μ^L(i)
    end if
  end if
  if Ind = 1 then
    if ΔP_H ≤ γ_stop and Count == n_fr then
      Ind = 0
      High_old = P_μ^H(i)
      Low_old = P_μ^L(i)
      Count = 0
      return
    else
      if Count < n_fr then
        Count = Count + 1
      else
        Initialize IMCRA variables as at the first frame
        for all frequency bins
      end if
    end if
  end if
end for
```

4. CROSS-PSD ESTIMATOR

In this section, we propose a novel scheme based on IMCRA to estimate the off-diagonal elements of the noise PSD matrix $\mathbf{P}(k,i)$ in (5). In this scheme, the noise component in each microphone signal is first estimated by means of a selected single channel speech enhancement algorithm which employs the estimated auto-PSD for the corresponding channel. Using the estimated noise components from different microphone pairs, the cross-PSDs can then be obtained by recursive smoothing as in IMCRA.

4.1. IMCRA Based Cross-PSD Estimator

We have been able to observe that the presence of speech components negatively impact the estimation of the noise cross-PSD when applying an IMCRA type of recursive smoother. On this basis, we propose to estimate the cross-PSD $P_{\mu,\nu}(k,i)$ in (4) by recursive smoothing of cross-periodograms derived from the estimated noise components in the corresponding microphone channels, instead of the observed noisy speech components.

Specifically, the proposed cross-PSD estimate, for a given pair of microphones with indices $\mu \neq \nu$, is obtained as

$$\begin{aligned}\hat{P}_{\mu,\nu}(k,i) &= \tilde{\alpha}_c(k,i)\hat{P}_{\mu,\nu}(k,i-1) \\ &\quad + (1-\tilde{\alpha}_c(k,i))\hat{N}_\mu(k,i)\hat{N}^*_\nu(k,i) \quad (12)\end{aligned}$$

where

$$\tilde{\alpha}_c(k,i) \triangleq \alpha_c + (1-\alpha_c)p(k,i) \quad (13)$$

is a time-varying frequency-dependent smoothing parameter with lower bound $0 < \alpha_c < 1$, and $\hat{N}_\mu(k,i)$ is the estimated noise component for frequency bin k and time frame i of the μth microphone signal.

The above recursive update is similar in nature to the IMCRA-based update (6)-(7) employed here to estimate the auto-PSD. The main difference lies in the use of the estimated noise components $\hat{N}_\mu(k,i)$, as opposed to the observed noisy speech components $Y_\mu(k,i)$, in forming the cross-periodogram terms. The removal of the speech components from the observations makes it possible to reduce the value of α_c, as compared to α in (7), which in turn is equivalent to the use of a shorter averaging window. Another difference with (6)-(7) is in the calculation of the smoothing parameter $\tilde{\alpha}_c(k,i)$, where we now use the maximum conditional speech presence probability over all the available microphone channels, that is:

$$p(k,i) = \max_\mu \{p_\mu(k,i)\}, \quad (14)$$

where $p_\mu(k,i)$ denotes the conditional speech presence probability computed as in IMCRA and the maximum is over all microphone channels. This approach tends to give slightly better estimates of the cross-PSD.

4.2. Noise Estimation

In the proposed algorithm, the estimated noise components $\hat{N}_\mu(k,i)$ are obtained by taking advantage of a selected SM speech enhancement algorithm applied separately to each one of the microphone signals.

Specifically, for a given microphone channel μ, the estimated noise component $\hat{N}_\mu(k,i)$ is computed as

$$\hat{N}_\mu(k,i) = Y_\mu(k,i) - \hat{S}_\mu(k,i) \quad (15)$$

where

$$\hat{S}_\mu(k,i) = G_\mu(k,i)Y_\mu(k,i) \quad (16)$$

denotes the enhanced speech STFT component and $G_\mu(k,i)$ is the corresponding enhancement gain, which can be calculated by any SM speech enhancement algorithm. In this paper, we use both the MMSE-based gain function from [10] and the OM-LSA gain function from [11] for this calculation, and compare the performance of the resulting noise PSD matrix estimators. In both cases, the proposed auto-PSD estimator $\hat{P}_{\mu\mu}(k,i)$ for microphone channel μ is employed in the calculation of the corresponding gain.

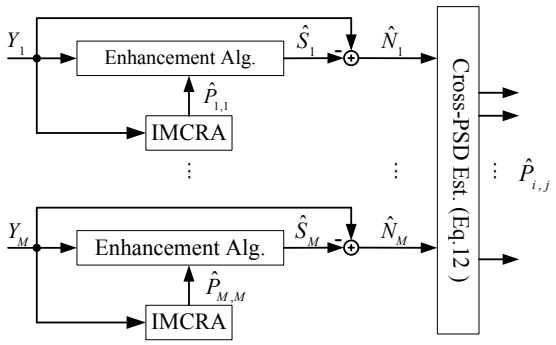

Fig. 1. Proposed cross-PSD estimator

5. RESULTS

In this section, we present the results of simulation experiments aimed at evaluating the performance of the proposed noise PSD matrix estimation algorithms.

5.1. Experimental Setup

We consider MA acquisition of a desired speech signal in the presence of noise in a rectangular room with dimensions $4 \times 5 \times 3$ (all units in meters). The image method [13] with refinement for non-integer delays is employed to emulate acoustic propagation between two points in the room. Two different acoustic environments are employed, that is: without reverberation and with moderate level of reverberation where the walls, ceiling and floor reflection coefficients are set to 0.70, 0.55 and 0.40, respectively. We use $M = 2$ microphones located 0.4 apart (horizontally) at positions $[1.8, 2.0, 1.25]$ and $[2.2, 2.0, 1.25]$, while the speech and noise sources are located at $[1.9, 1.5, 1.25]$ and $[3, 4, 2]$, respectively.

Six speech files from 3 male and 3 female speakers are used in the experiments. Each file is constructed by concatenating 10 short sentences from the same speaker without intervening pauses. The speech signals are degraded by various types of noise with SNR varying from -5 to 15dB in steps of 5dB. The noise files include a non-stationary white Gaussian noise (WGN) with sudden level increase, air conditioning (AC) fan noise and hallway noise (see Fig. 2 for additional information). All the signals are sampled at 16kHz while for the STFT analysis, we use a 512-point FFT, a hamming window, and an overlap of 256 samples.

These files are used to evaluate the quality of the newly proposed noise PSD matrix estimator. For auto-PSD estimation, we compare the performance of the modified IMCRA proposed in Section III to that of the conventional IMCRA from [3]. For the complete PSD matrix, with auto and cross-PSD estimation from Section III and IV, respectively, we consider two different versions of the proposed algorithm:

- ◇ *Mod-MMSE*: Modified IMCRA for auto-PSD with proposed cross-PSD based on MMSE gain from [10]

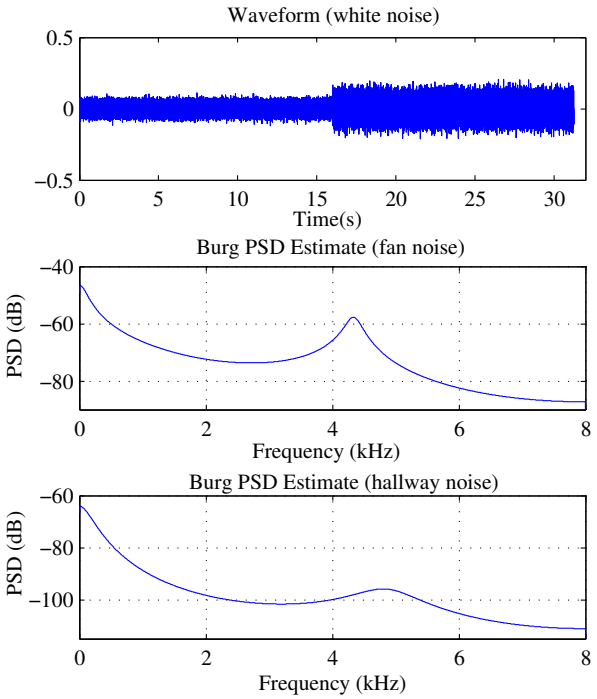

Fig. 2. Noise signals used in experiments. From top to bottom: non-stationary WGN, AC fan noise and hallway noise

- ◇ *Mod-OMLSA*: Modified IMCRA for auto-PSD with proposed cross-PSD based on OM-LSA gain from [11]

These are compared to two selected algorithms from the recent literature, namely:

- ◇ *Algo-H*: Noise PSD matrix estimator from [8];
- ◇ *Algo-F*: VAD-based estimator from [6].

Note that *Algo-H* requires *a priori* knowledge of the propagation vector $\mathbf{d}(k)$ between the speaker and the MA. Here, we use the exact $\mathbf{d}(k)$ derived from the room impulse responses, but in practice, this vector would need to be estimated.

5.2. Performance Measures

Several objective measures are employed to evaluate the performance of the proposed noise PSD matrix estimation algorithm. For the auto-PSD estimator, we use the log spectral distance (LSD) which is defined for the ith frame as

$$\mathrm{LSD}_{\mu}(i) = \sqrt{\frac{1}{L}\sum_{k=0}^{L-1}\left[10\log_{10}\frac{P_{\mu,\mu}(k,i)}{\hat{P}_{\mu,\mu}(k,i)}\right]^2} \qquad (17)$$

where $P_{\mu,\mu}(k,i)$ is the ideal noise auto-PSD (i.e., obtained from the noise-only file) and $\hat{P}_{\mu,\mu}(k,i)$ is the estimated one. For the complete noise PSD matrix estimator, including the cross-PSD estimator in Section 4.1, we resort to a so-called

Frobenius spectral distance, defined for the ith frame as

$$\text{FSD}(i) = \sqrt{\frac{1}{L}\sum_{k=0}^{L-1} \|\mathbf{P}(k,i) - \hat{\mathbf{P}}(k,i)\|_F^2} \qquad (18)$$

where $\|.\|_F$ denotes the Frobenius norm, $\mathbf{P}(k,i)$ is the ideal noise PSD matrix and $\hat{\mathbf{P}}(k,i)$ is the estimated one.

To evaluate the overall quality of the proposed noise PSD matrix estimator, we also consider its effect when used in combination with a MA speech enhancement algorithm based on the MVDR beamformer. The weight vector of this beamformer is given by [4]

$$\mathbf{w}(k) = \frac{\hat{\mathbf{P}}(k,i)^{-1}\mathbf{d}(k)}{\mathbf{d}^{\mathbf{H}}(k)\hat{\mathbf{P}}(k,i)^{-1}\mathbf{d}(k)} \qquad (19)$$

where here, the steering vector $\mathbf{d}(k)$ can be obtained from the synthesized room impulse responses. Using this weight vector, the MVDR beamformer output is computed as

$$\hat{S}(k,i) = \mathbf{w}^H(k)\mathbf{Y}(k,i) \qquad (20)$$

where $\mathbf{Y}(k,i) = [Y_1(k,i), \ldots, Y_M(k,i)]^T$ and $\hat{S}(k,i)$ denotes the enhanced speech at the beamformer output. Finally, we compute the PESQ-MOS [14] between the reconstructed enhanced and clean speech (in the time-domain) as an objective performance measure.

5.3. Results and Discussion

Experiment 1: In this experiment, we study the effect of a sudden increase in the background noise level on the performance of the proposed noise PSD matrix estimator. The noise waveform used for this experiment is shown in Fig. 2 (top), where the noise power is increased by about 6dB at time 16s. This waveform is added to a selected speech file so that the overall SNR=0dB (no reverberation).

We first compare the performance of the modified IMCRA proposed in Section 3.2 for auto-PSD estimation to that of the conventional IMCRA [3]. To this end, Fig. 3 shows the time evolution of the LSD (17) at a selected microphone for the two algorithms. From the results, it can be seen that the conventional IMCRA takes around 260 frames to recover from the abrupt change, whereas the modified IMCRA converges much faster. We generally find that the performance of the modified IMCRA in tracking the noise auto-PSD is superior (e.g. in the case of a sudden noise increase), or at least similar to that of the conventional one.

Next, we evaluate the overall performance of the proposed noise PSD matrix estimator. Fig. 4 shows the time evolution of the FSD (18) for the proposed *Mod-MMSE* and *Algo-H* algorithms under the same scenario of a sudden noise change as in Fig. 3. Again, it can be seen that our proposed algorithm leads to a better performance, not only in recovering from the sudden noise change, but also in maintaining a lower level of residual FSD during the stationary portions of the noise background before and after the sudden change.

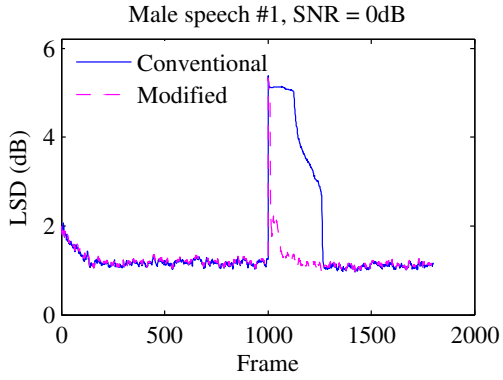

Fig. 3. LSD comparison between modified and conventional IMCRA algorithms for auto-PSD estimation

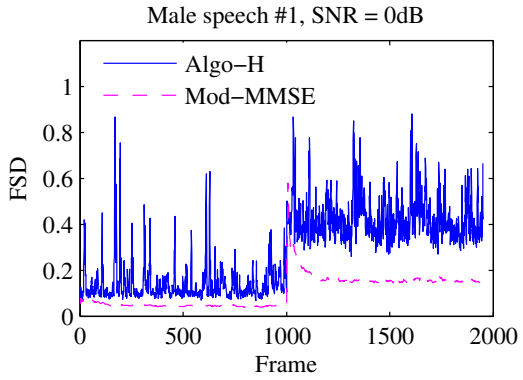

Fig. 4. FSD comparison between proposed noise PSD matrix estimation and algorithm from [8]

Experiment 2: In this experiment, we study the performance of the proposed noise PSD matrix estimator when used in combination with the MVDR beamformer (19)-(20). For each one of the four algorithms listed in Section 5.1, the PESQ-MOS of the enhanced speech at the beamformer output is calculated and averaged over the six different speakers. This is repeated for different noise types and SNR values.

Table 1 lists the PESQ-MOS obtained in this way with the four noise PSD matrix estimators in the absence of reverberation. In all cases, the two versions of the proposed algorithm, i.e. *Mod-MMSE* and *Mod-OMLSA*, achieve the best performance. Furthermore, the use of the MMSE gain function from [10] in the noise estimation (15)-(16) leads to better enhancement results, suggesting that this method is more appropriate for use in connection with the proposed noise cross-PSD estimator.

Table 2 lists the PESQ-MOS of the four noise PSD ma-

Table 1. PESQ-MOS of MVDR Beamformer using Different Noise PSD Matrix Estimators (no reverberation)

Noise type	Estimator	SNR (dB)				
		-5	0	5	10	15
non-stat WGN	Mod-MMSE	1.92	2.34	2.50	2.66	2.80
	Mod-OMLSA	1.41	1.81	1.99	2.25	2.51
	Algo-H	1.43	1.76	2.03	2.22	2.37
	Algo-F	0.99	1.17	1.45	1.72	1.87
fan noise	Mod-MMSE	2.27	2.57	2.67	2.89	3.02
	Mod-OMLSA	1.83	2.07	2.23	2.55	2.77
	Algo-H	1.76	2.02	2.20	2.37	2.51
	Algo-F	1.19	1.25	1.53	1.80	2.05
hallway noise	Mod-MMSE	2.35	2.67	2.78	3.00	3.08
	Mod-OMLSA	2.05	2.34	2.50	2.75	2.90
	Algo-H	1.87	2.07	2.23	2.37	2.52
	Algo-F	1.19	1.36	1.58	1.81	2.00

Table 2. PESQ-MOS of MVDR Beamformer using Different Noise PSD Matrix Estimators (with reverberation)

Noise type	Estimator	SNR (dB)				
		-5	0	5	10	15
non-stat WGN	Mod-MMSE	1.62	1.96	2.13	2.25	2.31
	Mod-OMLSA	1.29	1.63	1.88	2.06	2.17
	Algo-H	1.03	1.29	1.59	1.79	1.82
	Algo-F	0.86	1.00	1.25	1.48	1.65
fan noise	Mod-MMSE	2.06	2.24	2.31	2.38	2.39
	Mod-OMLSA	1.80	1.97	2.13	2.23	2.29
	Algo-H	1.45	1.74	1.88	1.94	1.96
	Algo-F	0.98	1.25	1.48	1.64	1.74
hallway noise	Mod-MMSE	1.95	2.17	2.28	2.37	2.39
	Mod-OMLSA	1.78	2.00	2.17	2.27	2.30
	Algo-H	1.58	1.80	1.93	1.98	2.00
	Algo-F	1.11	1.30	1.48	1.61	1.71

trix estimators, but this time in the presence of reverberation. Comparing corresponding entries in Table 1 and 2, we note that reverberation degrades the speech enhancement performance in all cases, with a noticeable reduction in PESQ-MOS. Nevertheless, the same conclusions as above can be made regarding the relative performance of the four algorithms, with the proposed noise PSD matrix estimators *Mod-MMSE* and *Mod-OMLSA* giving the best results by a significant margin.

6. CONCLUSIONS

In this paper, we presented a novel method to estimate the noise PSD matrix for MA systems, which consists of two parts. For the auto-PSD estimation, we introduced a modification to IMCRA where a special level detector is employed to improve the tracking of non-stationary noise backgrounds. In comparison to the original IMCRA in [3], the proposed algorithm converges much faster when the noise level is suddenly increased. For the cross-PSD estimation, we proposed to calculate a smoothed cross-periodogram by using estimated noise components instead of the noisy speech signals received from the microphones. The noise estimates can be obtained as residuals after the application of a selected SM speech enhancement algorithm on the individual microphone signals. Simulation results showed the effectiveness of our proposed approach in estimating the noise PSD matrix, and its robustness against reverberation when applied to a speech enhancement system based on MVDR beamforming.

7. REFERENCES

[1] P. L. Loizou, *Speech Enhancement: Theory and Practice,* CRC, 2011.

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

[3] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Trans. on Speech and Audio Processing*, vol. 11, pp. 466–475, May 2003.

[4] M. Brandstein and D. Ward (Eds.), *Microphone Arrays: Signal Processing Techniques and Applications*, Springer-Verlag, 2008.

[5] X. Zhang and Y. Jia, "A soft decision based noise cross power spectral density estimation for two-microphone speech enhancement systems," in *Proc. Intl. Conf. on Acoustics, Speech, and Signal Processing* (Philadelphia, PA), vol. 1, pp. 813–816, March 2005.

[6] J. Freudenberger, S. Stenzel, and B. Venditti, "A noise PSD and cross-PSD estimation method for two-microphone speech enhancement systems," in *Proc. IEEE Workshop on Statistical Signal Processing*, pp. 709–712, Sept. 2009.

[7] A. H. Kamkar-Parsi, and M. Bouchard, "Improved noise power spectral density estimation for binaural hearing aids operating in a diffuse noise field environment," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 17, pp. 521–533, May 2009.

[8] R. C. Hendriks, and T. Gerkmann, "Noise correlation matrix estimation for multi-microphone speech enhancement," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, pp. 223–233, Jan. 2012.

[9] N. Fan, J. Rosca, and R. Balan, "Speech noise estimation using enhanced minima controlled recursive averaging," in *Proc. ICASSP* (Honolulu, USA), vol. IV, pp. 581-584, May 2007.

[10] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized gamma priors," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, pp. 1741–1752, Aug. 2007.

[11] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, pp. 2403-2418, 2001.

[12] J. Taghia, N. Mohammadiha, J. Sang, V. Bouse and R. Martin, "An evaluation of noise power spectral density estimation algorithms in adverse acoustic environments," in *Proc. ICASSP* (Prague, Czech), pp. 4640–4643, May 2011.

[13] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoustic Society of America*, vol. 65, no. 4 pp. 943–950, Apr., 1979.

[14] ITU-T, *Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-end Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs*, ITU-T Rec. P.862, Nov. 2005.