Engineering Journal issn: 0125-8281

Article

A Multi-Channel Noise Estimator Based on Improved Minima Controlled Recursive Averaging for Speech Enhancement

Nisachon Tangsangiumvisai

Multimedia Data Analytics and Processing Research Unit, Department of Electrical Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok 10330, Thailand
E-mail: Nisachon.T@chula.ac.th (Corresponding author)

Abstract. This article introduces an extension of the improved minima-controlled recursive averaging noise estimation from single to multi-channel speech enhancement systems. With the spatial information of microphone array signals being fully exploited, more accurate estimate of the noise spectrum can be obtained over the single-channel counterpart. Computer simulation demonstrates superior performance of the proposed noise estimator in terms of noise tracking performance and noise estimation error. Furthermore, the use of the proposed technique with the multi-channel Wiener filter yields improved signal-to-noise ratio and speech distortion.

Keywords: Noise estimation, speech enhancement, multi-channel, microphone array, signal-to-noise ratio.

ENGINEERING JOURNAL Volume 27 Issue 11
Received 26 September 2023
Accepted 20 November 2023
Published 30 November 2023
Online at http://www.engj.org/
DOI:10.4186/ej.2023.27.11.99

DOI:10.4186/ej.2023.27.11.99

1. Introduction

Modern voice communication devices such as smartphones, hands-free car kits, tablets, or desktop computers, are usually well-equipped with microphones and loudspeakers. Furthermore, electrical appliances in smart homes and smart factories nowadays have been arranged to accommodate voice-controlled applications, such as virtual meeting, etc. In addition, assistive listening devices, such as hearing aids for hearing-impaired people, usually require multichannel processing. However, the problem of ambient noise is typically unavoidable and is always included in the microphone signals of these assistive devices. Speech enhancement techniques are therefore necessary to improve the quality of speech signals by removing the effect of the additive noise.

Various speech enhancement techniques based on the frequency-domain processing have been employed to eliminate the additive noise. The spectral subtraction (SS) method, which can be realized as spectral suppression, is one of the well-known speech enhancement approaches. Basically, the SS method and its modified versions obtain the enhanced speech spectrum by subtracting the estimated noise spectrum from the noisy speech spectrum [1]−[3]. Depending on the efficiency of voice activity detectors (VAD), the noise spectrum is usually estimated during speech-absence frames. It is obvious that the noise estimator plays an important role for frequency-domain speech enhancement algorithms. A number of noise estimation techniques have been proposed for single-microphone or single-channel speech enhancement [4]−[9]. Some of these noise estimation techniques update the noise spectrum during speech-absence frames, while several of them continuously update the noise spectrum for both speech-absence and speech-presence frames. The accuracy of the noise spectral estimate affects completely the overall performance of the speech enhancement algorithms; i.e., noise overestimation results in suppression of the speech spectral components and introduces noticeable speech distortion. On the other hand, noise underestimation results in the residual noise in the enhanced speech signal, which causes poor speech quality.

Conversely to the single-channel speech enhancement techniques, multi-channel approaches tend to be more effective with the use of spatial-temporal information. In fact, microphone arrays have been exploited extensively for speech enhancement in most modern voice communication systems in order to deal with the problem of room reverberation and ambient noise [10]−[16]. Basically, a beamformer adjusts the values of gain and phase differently for different microphones within the system, depending on the *a priori* knowledge about the location of the target source, so that the desired speech signal is emphasized, whereas the ambient noise and interference sources are diminished. Various types of beamforming techniques can be chosen based on the constraints that best fit specific applications. The delay-and-sum (DAS) beamformer is one of the most common types where the signals arriving at each microphone have different delays. The sum of these microphone signals, which are scaled differently, is maximum when the steering direction is according to the desired source direction. Linearly constrained minimum variance (LCMV) beamformer is chosen to reduce the effect of reverberant and ambient noise provided that general acoustic transfer functions (ATFs) between the source and the microphones are known [13], [14]. Minimum variance distortionless response (MVDR), on the other hand, becomes more popular than the normal LCMV beamformer because of its special constraint [15]. Then, the generalized sidelobe canceler (GSC), which is an unconstrained form of the LCMV beamformer, was introduced as an alternative to extract the desired speech signal from its noisy microphone signals [16]. Another type of multichannel speech enhancement algorithms based on the optimal multi-dimensional filter is known as multi-channel Wiener filtering (MWF) [17]−[21]. A general parameterized expression for MWF is derived in [17]. A multimicrophone optimal filter is designed based on a single-microphone subspace-based for white and colored additive noise [18]. In [19], a multi-microphone speech enhancement algorithm was introduced based on generalized singular value decomposition (GSVD) of speech and noise data matrices. A frequency-domain criteria for speech distortion weighted multi-channel Wiener filter (SDW-MWF) was introduced in [20]. Then, the frequency-domain spatial-prediction method was developed and yielded better performance for the estimation of the specech covariance matrix, as compared to the conventional MWF [21].

The multi-channel noise estimation techniques are therefore essential for these multi-channel speech enhancement algorithms so that all spatial-temporal information from the microphone array is utilized to give a better noise estimate. Most of the existing noise estimation techniques are, however, formulated for single-channel communication systems [4]−[9]. The noise power spectrum can be estimated based on minimum statistics (MS) without any use of VAD; i.e., it is assumed that the noisy speech power spectrum usually decays to the noise power level, hence the noise spectral estimate in each frequency band can be obtained by tracking the spectral minimum values in that particular frequency-band [4]. Subsequently, this noise tracking technique was further employed in the minima controlled and recursive averaging (MCRA) approach and its improved version (IMCRA), especially in adverse environments such as low input signal-to-noise ratio (SNR) and non-stationary noises [5], [6]. Specifically, the MCRA and IMCRA approaches introduce speech absence probability (SAP) and speech presence probability (SPP) for continu-

ous noise estimation, even during frames with weak speech activity [6]. In addition, the IMCRA approach employs a smoothing parameter and bias compensation factor, both in time- and frequency-domain to improve the accuracy of the noise spectrum estimation. Alternatively, noise estimation techniques based on quantile-based have been introduced without prior knowledge of speech and noise distributions [7]−[9]. These quantile-based noise estimation techniques, however, encounter the problem of high computational complexity for calculating and storing the quantile information.

So far, only a few multi-channel noise estimation techniques have been introduced for multi-channel speech enhancement [22]−[24]. In [22], a multi-channel noise estimation technique was introduced based on the MCRA technique in [5] in order to obtain an optimal *a priori* SAP estimation. However, one of the major limitations of this technique in [22] is the requirement of initial noise-only frames in order to initialise the noise covariance matrix for subsequent noise power spectrum estimation, which may not be readily available or accurately determined in practice. In fact, a sample average of a fixed duration of periodograms is subsequently used to approximate the noise covariance matrix. As a result, this kind of estimation of the noise covariance matrix is subject to inaccuracy, particularly during the occurrence of any abrupt SNR changes. In [23], speech and noise covariance matrices are estimated based on eigenvalue decomposition. Therefore, this technique entails high computational complexity, especially during the eigenvalue decomposition process. In [24], the multi-channel SPP is employed for estimating speech statistics. Nonetheless, the SAP employed in this algorithm is chosen as a constant and kept fixed for all frames and frequencies. The use of a fixed SAP does not allow the algorithm to track efficiently the non-stationary characteristics of the ambient noise signals. Thus, accurate noise spectrum estimate cannot be obtained.

Therefore, it is proposed in this article to formulate the multi-channel noise estimator, based on the well-established single-channel IMCRA noise estimation technique [6]. The proposed multi-channel noise estimator employs the minimum tracking performance of the IMCRA technique and fully exploit the spatial information of multi-channel system so as to compute more accurate SAP and SPP estimates. Hence, the proposed noise estimation technique is supposed to estimate the noise covariance matrix more accurately, particularly for multi-channel speech enhancement applications, makes it suitable for highly non-stationary noise typically encountered in modern devices.

This article is organized as follows. The signal model is presented in Section 2, followed by fundamental principles and the detailed formulation of the proposed multi-channel noise spectral estimator, based on the IMCRA algorithm in Section 3. Then, simulation results are presented to illustrate the advantages of the proposed noise estimator, as compared to the investigated single-channel noise estimation technique in Section 4. Finally, conclusions are given in Section 5.

2. Signal Model

An M-element microphone array is considered in noisy and reverberant environment. It is assumed that there is a single target speech source, $s(n)$, which is assumed to be uncorrelated to the ambient noise in each microphone, $v_m(n)$, for $m = 1, 2, \ldots, M$. The reverberant speech signal for the m-th microphone, $x_m(n)$, is obtained by convolving room impulse response (RIR), $d_m(n)$, with the clean speech signal, $s(n)$; $x_m(n) = s(n) * d_m(n)$, where $< * >$ denotes the convolution operator. Each microphone signal is then analyzed via the short-time Fourier transform (STFT), and the relationship in the frequency-domain becomes

$$\mathbf{y}(k,l) = \mathbf{x}(k,l) + \mathbf{v}(k,l) \tag{1}$$

where $\mathbf{y}(k,l) = [Y_1(k,l), Y_2(k,l), \ldots, Y_M(k,l)]^T$ is the vector containing M microphone spectra, the parameter $k = 0, 1, \ldots, K-1$ denotes the frequency-bin index for K-point STFT, and $l = 0, 1, \ldots, L-1$ is the frame index when L is the total number of frames. Similarly, $\mathbf{x}(k,l) = [X_1(k,l), X_2(k,l), \ldots, X_M(k,l)]^T$ is the vector containing M reverberant speech spectra, and $\mathbf{v}(k,l) = [V_1(k,l), V_2(k,l), \ldots, V_M(k,l)]^T$ is the vector containing M noise spectra. The reverberant speech spectral vector is given by $\mathbf{x}(k,l) = \mathbf{d}(k)S(k,l)$, where $\mathbf{d}(k) = [D_1(k), D_2(k), \ldots, D_M(k)]^T$ is the so-called acoustic transfer function (ATF) vector between the speech source to all M microphones, and $S(k,l)$ is the STFT of the speech signal, $s(n)$.

For speech enhancement, it is desirable to obtain the estimated clean speech spectrum $\hat{S}(k,l)$. To achieve this purpose, the noise spectrum has to be estimated first. Based on the hypothesis testing where $H_0(k,l)$ denote the speech absence and $H_1(k,l)$ denote the speech presence, the conditional probability density function (PDF) of the microphone spectra are given by

$$f(\mathbf{y}(k,l)|H_0(k,l)) = \frac{1}{\pi^M \det\big(\mathbf{R_{vv}}(k,l)\big)} \exp\{-\mathbf{y}^H(k,l)\mathbf{R}_{\mathbf{vv}}^{-1}(k,l)\mathbf{y}(k,l)\} \tag{2}$$

and

$$f(\mathbf{y}(k,l)|H_1(k,l)) = \frac{1}{\pi^M \det\Big(\mathbf{R_{xx}}(k,l) + \mathbf{R_{vv}}(k,l)\Big)} \cdot \exp\{-\mathbf{y}^H(k,l)\big(\mathbf{R_{xx}}(k,l) + \mathbf{R_{vv}}(k,l)\big)^{-1}\mathbf{y}(k,l)\} \tag{3}$$

where the clean speech spectrum, $\mathbf{x}(k,l)$, and noise spectrum, $\mathbf{v}(k,l)$, in eq.(1) are both assumed to

DOI:10.4186/ej.2023.27.11.99

be complex Gaussian distribution with zero mean. $\mathbf{R}_{\mathbf{xx}}(k,l) = E\{\mathbf{x}(k,l)\mathbf{x}^H(k,l)\}$ and $\mathbf{R}_{\mathbf{vv}}(k,l) = E\{\mathbf{v}(k,l)\mathbf{v}^H(k,l)\}$ are the speech and noise covariance matrices, respectively, and $E\{\cdot\}$ denotes the expectation operator. By assuming that the clean speech and noise signals are uncorrelated to each other, we obtain that the multi-channel covariance matrix of the noisy speech spectrum, $\mathbf{R}_{\mathbf{yy}}(k,l) = E\{\mathbf{y}(k,l)\mathbf{y}^H(k,l)\}$, is given by

$$\mathbf{R}_{\mathbf{yy}}(k,l) = \mathbf{R}_{\mathbf{xx}}(k,l) + \mathbf{R}_{\mathbf{vv}}(k,l). \tag{4}$$

3. The Proposed Multi-channel Noise Estimation Approach

In this section, the proposed multi-channel noise estimation approach is formulated based on the single-channel IMCRA noise estimation technique [6], and will be hereafter referred to as the MC-IMCRA technique. The proposed technique makes use of the minimum tracking operation of the IMCRA technique while fully exploiting the spatial information of multi-channel systems in order to compute more accurate SAP and SPP estimates. Hence, the proposed multi-channel noise estimator is able to estimate the noise covariance matrix more accurately.

Let the multi-channel *a priori* signal-to-noise ratio (SNR) and the multi-channel instantaneous *a posteriori* SNR be defined as follows:

$$\xi(k,l) = \mathrm{tr}\left[\mathbf{R}_{\mathbf{vv}}^{-1}(k,l)\mathbf{R}_{\mathbf{xx}}(k,l)\right] \tag{5}$$

$$\gamma(k,l) = \mathbf{y}^H(k,l)\mathbf{R}_{\mathbf{vv}}^{-1}(k,l)\mathbf{y}(k,l) \tag{6}$$

where tr[·] denote the trace of a matrix. Given the *a priori* speech absence probability (SAP) as

$$q(k,l) = P(H_0(k,l)), \tag{7}$$

the conditional speech presence probability (SPP) is obtained, based on the Bayes' theorem, as

$$p(k,l) = P(H_1(k,l)|\gamma(k,l)), \tag{8}$$

which becomes [6]

$$p(k,l) = 1 + \frac{q(k,l)}{1-q(k,l)}\left[1+\xi(k,l)\mathrm{e}^{-\frac{\gamma(k,l)\xi(k,l)}{1+\xi(k,l)}}\right]^{-1}. \tag{9}$$

Based on the hypotheses of speech absence, $H_0(k,l)$, and speech presence, $H_1(k,l)$, the estimated noise covariance matrix, $\hat{\mathbf{R}}_{\mathbf{vv}}(k,l)$, can be obtained by recursive update, particularly during speech absence frames, as follows:

$$\begin{aligned}\hat{\mathbf{R}}_{\mathbf{vv}}(k,l+1) =& \alpha_v(k,l)\hat{\mathbf{R}}_{\mathbf{vv}}(k,l) \\ &+ \beta_1\Big(1-\alpha_v(k,l)\Big)\mathbf{y}(k,l)\mathbf{y}^H(k,l).\end{aligned} \tag{10}$$

The forgetting factor $\alpha_v(k,l)$ is employed to track the noise spectrum for all frequency bins and all frames, based upon the SPP,

$$\alpha_v(k,l) = \tilde{\alpha}_v + (1-\tilde{\alpha}_v)\,p(k,l) \tag{11}$$

and $0 \le \tilde{\alpha}_v < 1$ is normally chosen. The function of the SPP is to bias the update of the forgetting factor $\alpha_v(k,l)$ towards high values in order to avoid speech distortion when $p(k,l)$ is close to unity. On the other hand, the SPP is to bias the update of $\alpha_v(k,l)$ towards lower values for noise spectral estimate when $p(k,l)$ approaches zero. Similar to the single-channel case, the bias compensation factor $\beta_1 = \mathrm{tr}(\mathbf{R}_{\mathbf{vv}}(k,l))/E\{\hat{\mathbf{R}}_{\mathbf{vv}}(k,l)\}$ is also introduced for updates during speech absent frames.

Since the *a priori* SAP, $q(k,l)$, is unknown, an estimator based on the tracking of minima values of smoothed multi-channel noisy power spectrum is computed. The multi-channel extension of the IMCRA noise estimation technique comprises two iterations, $j = 1, 2$. The computation required for both iterations are summarised as follows. First, the smoothing operation is carried out both in the frequency-domain and the time-domain. The frequency smoothing of the noisy power spectrum is taken for each frame as

$$\mathbf{S}_{\mathbf{y},f}^{(j)}(k,l) = \begin{cases} \dfrac{\sum_{i=-N_b}^{N_b} b(i)I^{(j)}(k-i,l)\mathbf{y}(k-i,l)\mathbf{y}^H(k-i,l)}{\sum_{i=-N_b}^{N_b} b(i)I^{(j)}(k-i,l)}, \\ \qquad \text{if } \sum_{i=-N_b}^{N_b} I^{(j)}(k-i,l) \neq 0 \\ \mathbf{S}_{\mathbf{y}}^{(j)}(k,l-1), \qquad \text{otherwise} \end{cases} \tag{12}$$

where $b(i)$ is a window function of length $2N_b + 1$. Next, the time smoothing is performed on the speech spectral components obtained from the frequency smoothing, by using a first-order recursive averaging,

$$\mathbf{S}_{\mathbf{y}}^{(j)}(k,l) = \alpha_s\mathbf{S}_{\mathbf{y}}^{(j)}(k,l-1) + (1-\alpha_s)\,\mathbf{S}_{\mathbf{y},f}^{(j)}(k,l) \tag{13}$$

where $0 \le \tilde{\alpha}_s < 1$ is another forgetting factor. Afterwards, the minimum tracking operation is performed. The minima value of $\mathbf{S}_{\mathbf{y}}^{(j)}(k,l)$ is found for each frequency bin over windowed frames of length N_w.

$$\mathbf{S}_{\mathbf{y},\min}^{(j)}(k,l) = \min\left\{\mathbf{S}_{\mathbf{y}}^{(j)}(k,\tilde{l}) \,\middle|\, l - N_w + 1 < \tilde{l} < l\right\}. \tag{14}$$

Note that, the window of N_w samples is divided into U sub-windows of V samples, where $U \times V = N_w$, as implemented in [6].

Table 1. The proposed multi-channel noise estimation algorithm (MC-IMCRA) for speech enhancement.

1. Initialization of all relevant parameters
- $\hat{\mathbf{R}}_{\mathbf{vv}}(k,0) = \mathbf{y}(k,0)\mathbf{y}^H(k,0)$
- $\gamma(k,0) = 1$
- $\mathbf{S}_{\mathbf{y}}^{(1)}(k,0) = \sum_{i=-N_b}^{N_b} b(i)\mathbf{y}(k-i,0)\mathbf{y}^H(k-i,0)$
- $\mathbf{S}_{\mathbf{y}}^{(2)}(k,0) = \mathbf{S}_{\mathbf{y}}^{(1)}(k,0)$
- $\mathbf{S}_{\mathbf{y},\min}^{(1)}(k,0) = \mathbf{S}_{\mathbf{y},\min}^{(2)}(k,0) = \mathbf{S}_{\mathbf{y}}^{(1)}(k,0)$

2. The MC-IMCRA algorithm

2.1. Run for all frames l and for all frequency bins k, and compute the following parameters.
- $\hat{\xi}(k,l)$ by using eq.(26)
- $\hat{\gamma}(k,l)$ by using eq.(6) with $\hat{\mathbf{R}}_{\mathbf{vv}}(k,l)$

2.2. Repeat the following computations for $j = 1$ and $j = 2$.
- $\mathbf{S}_{\mathbf{y}}^{(j)}(k,l)$ by using eq.(13)
- $\mathbf{S}_{\mathbf{y},\min}^{(j)}(k,l)$ by using eq.(14)
- $\gamma_{\min}^{(j)}(k,l)$ by using eq.(15)
- $\zeta^{(j)}(k,l)$ by using eq.(16)
- $I^{(j)}(k,l)$ by using eq.(19)

3. Compute the estimated a priori SAP, SPP and the estimated noise covariance matrix.
- $\hat{q}(k,l)$ by using eq.(21)
- $\hat{p}(k,l)$ by using eq.(9) with $\hat{\xi}(k,l)$, $\hat{q}(k,l)$
- $\hat{\mathbf{R}}_{\mathbf{vv}}(k,l+1)$ by using eq.(10), eq.(11)

Subsequently, the following two parameters are computed, as given by

$$\gamma_{\min}^{(j)}(k,l) = \frac{1}{\beta_2}\left(\mathbf{y}^H(k,l)\big(\mathbf{S}_{\mathbf{y},\min}^{(j)}(k,l)\big)^{-1}\mathbf{y}(k,l)\right) \tag{15}$$

$$\zeta^{(j)}(k,l) = \frac{1}{\beta_2}\left(\big(\mathbf{S}_{\mathbf{y},\min}^{(j)}(k,l)\big)^{-1}\mathbf{S}_{\mathbf{y}}^{(j)}(k,l)\right) \tag{16}$$

where β_2 is a bias factor. In the speech absence case, the PDF of both $\gamma_{\min}^{(j)}(k,l)$ and $\zeta^{(j)}(k,l)$ can be modelled as the exponential and chi-square distributions, respectively. The thresholds γ_0 and ζ_0 are chosen so that the following conditions are satisfied for a small constant ϵ.

$$P\big(\gamma_{\min}(k,l) \geq \gamma_0 \big| H_0(k,l)\big) < \epsilon \tag{17}$$

$$P\big(\zeta(k,l) \geq \zeta_0 \big| H_0(k,l)\big) < \epsilon. \tag{18}$$

In the first iteration, $j = 1$, the frequency-smoothing operation in eq.(12) is determined with the decision function is set to be one at all frames and all frequency bins, i.e., $I^{(1)}(k,l) = 1$. Then, the time-smoothing operation in eq.(13) and eq.(14) are computed for $\mathbf{S}_{\mathbf{y}}^{(1)}(k,l)$ and $\mathbf{S}_{\mathbf{y},\min}^{(1)}(k,l)$. These values are therefore used for the calculation of $\gamma_{\min}^{(1)}(k,l)$ and $\zeta^{(1)}(k,l)$ in eq.(15), and eq.(16), and will be used for determining the decision function in the second iteration.

In the second iteration, $j = 2$, the frequency-smoothing operation in eq.(12) is computed using the following decision function:

$$I^{(2)}(k,l) = \begin{cases} 1, & \text{if} \gamma_{\min}^{(1)}(k,l) < \gamma_0 \text{ and } \zeta^{(1)}(k,l) < \zeta_0 \\ 0, & \text{otherwise.} \end{cases} \tag{19}$$

Note that, the decision function is defined to be one, $I^{(2)}(k,l) = 1$, when it is speech-absence frame, or zero, $I^{(2)}(k,l) = 0$, when it is speech-presence frame. With the above equation, the frequency smoothing operation in eq.(12) using $I^{(2)}(k,l)$ can remove most of the speech spectral components so that the noise spectrum can be estimated more accurately. Then, the time-smoothing operation is computed for $\mathbf{S}_{\mathbf{y}}^{(2)}(k,l)$ and $\mathbf{S}_{\mathbf{y},\min}^{(2)}(k,l)$ by employing eq.(13) and eq.(14). These values are then used for the calculation of $\gamma_{\min}^{(2)}(k,l)$ and $\zeta^{(2)}(k,l)$ in eq.(15), and eq.(16), respectively.

Finally, the *a priori* SAP can be estimated in two steps as follows. The local *a priori* SAP is firstly obtained, based on [25]:

$$\hat{q}_{\text{local}}(k,l) = \begin{cases} 1, & \text{if} \gamma_{\min}^{(2)}(k,l) \leq M \text{ and } \zeta^{(2)}(k,l) < \zeta_0 \\ \frac{\gamma_1 - \gamma_{\min}^{(2)}(k,l)}{\gamma_1 - M}, & \text{if } M < \gamma_{\min}^{(2)}(k,l) < \gamma_1 \\ & \text{and } \zeta^{(2)}(k,l) < \zeta_0 \\ 0, & \text{otherwise.} \end{cases} \tag{20}$$

From eq.(20), it can be seen that when $\gamma_{\min}^{(2)}(k,l)$ and $\zeta^{(2)}(k,l)$ are both less than the given thresholds, $\hat{q}(k,l) = 1$; i.e., the *a priori* SAP estimator decides that it is speech absence for the l-th frame and for the k–th frequency bin. On the other hand, when $\gamma_{\min}^{(2)}(k,l)$ and $\zeta^{(2)}(k,l)$ are both larger than the thresholds, $\hat{q}(k,l) = 0$; i.e., the *a priori* SAP estimator decides that it is speech presence case. Unlike the usual hard decision in conventional VAD, it can be seen that eq.(20) provides a soft decision between speech absence and speech presence conditions. Then, for the second step, the speech absence probability is finally computed, based on the local and frame-wise *a priori* SAP estimators.

$$\hat{q}(k,l) = \hat{q}_{\text{local}}(k,l) \cdot \hat{q}_{\text{frame}}(k,l) \tag{21}$$

The frame-wise *a priori* SAP estimator is defined as follows:

$$\hat{q}_{\text{frame}}(l) = \frac{1}{K}\sum_{k=0}^{K-1} \zeta(k,l) \tag{22}$$

The proposed multi-channel noise estimation algorithm can therefore be summarized in Table1.

DOI:10.4186/ej.2023.27.11.99

4. Multi-channel Speech Enhancement Algorithm

The multi-channel speech enhancement algorithm employed to verify the proposed MC-IMCRA technique is the SDW-MWF algorithm [20], and will be referred in short as MWF. The MWF algorithm is designed to suppress the spectral components of noisy speech spectrum when the SNR is low. On the other hand, when the SNR is high, the MWF will preserve the noisy speech spectral components. The spectral gain function $\mathbf{g}_W(k,l)$ of the MWF is given by

$$\mathbf{g}_W(k,l) = \frac{\frac{1}{M}\mathrm{tr}[\mathbf{R_{xx}}(k,l)]\mathbf{R_{vv}^{-1}}(k,l)\mathbf{d}(k)}{\mu + \left(\frac{1}{M}\mathrm{tr}[\mathbf{R_{xx}}(k,l)]\mathbf{d}^H(k)\mathbf{R_{vv}^{-1}}(k,l)\mathbf{d}(k)\right)} \tag{23}$$

where $0 < \mu < \infty$ is a tradeoff parameter controlling the noise reduction performance and the speech distortion level, and $\mathbf{d}(k)$ is ATF vector between the speech source to all M microphones [20]. The estimated covariance matrix of the noisy speech spectrum is obtained as

$$\hat{\mathbf{R}}_{\mathbf{yy}}(k,l) = \alpha\hat{\mathbf{R}}_{\mathbf{yy}}(k,l-1) + (1-\alpha)\mathbf{y}(k,l)\mathbf{y}^H(k,l) \tag{24}$$

and $0 < \alpha < 1$ is another forgetting factor. The estimated speech covariance matrix is obtained as

$$\hat{\mathbf{R}}_{\mathbf{xx}}(k,l) = \hat{\mathbf{R}}_{\mathbf{yy}}(k,l) - \hat{\mathbf{R}}_{\mathbf{vv}}(k,l). \tag{25}$$

The estimated *a priori* SNR for multi-channel systems is obtained by using the estimated speech and noise covariance matrices, as follows:

$$\hat{\xi}(k,l) = \mathrm{tr}\left[\hat{\mathbf{R}}_{\mathbf{vv}}^{-1}(k,l)\hat{\mathbf{R}}_{\mathbf{xx}}(k,l)\right] \tag{26}$$

Subsequently, the estimated noise covariance matrix is updated for the next frame $l+1$ by using eq.(9) – eq.(11). Therefore, the enhanced speech signal in the time-domain $\hat{s}(n)$ is obtained by taking the inverse STFT of $\hat{S}(k,l)$, where

$$\hat{S}(k,l) = \mathbf{g}_W^H(k,l)\mathbf{y}(k,l) \tag{27}$$

5. Simulation Sesults

In this section, computer simulations were carried out to observe the performance of the proposed MC-IMCRA noise estimation technique, as compared to the conventional IMCRA one. Then, the MWF algorithm [20] utilising the proposed MC-IMCRA technique for multi-channel speech enhancement (MWF+MC-IMCRA) was compared with the WF algorithm [28], which employed the IMCRA technique [6], for single-channel speech enhancement (WF+IMCRA).

5.1. Simulation Setup

A multi-channel speech enhancement system (MWF+MC-IMCRA) for an 8-channel microphone array was investigated. A 15-second clean speech signal, selected from various speech sentences in the IEEE database [29] at the sampling frequency of 8 kHz, was convolved with 8-channel RIRs from the hearing-aid head-related room impulse response (HRIR) database [30]. A zoom-in plot of one of the RIRs is given in Fig. 1. These multi-channel reverberant speech signals were then corrupted by the recorded 8-channel babble noises in a cafeteria environment [30] to represent the observed highly non-stationary noisy and reverberant speech signals from the microphone array. A sketch of the cafeteria environment is given in Fig. 2. The speaker was positioned in front of the dummy head wearing hear-aid devices on both side of its ears. Each side contains two types of the hearing devices. The one inside the ear canal has one microphone and the other one behind the ear has three microphones, thereby forming an eight-channel microphone array.

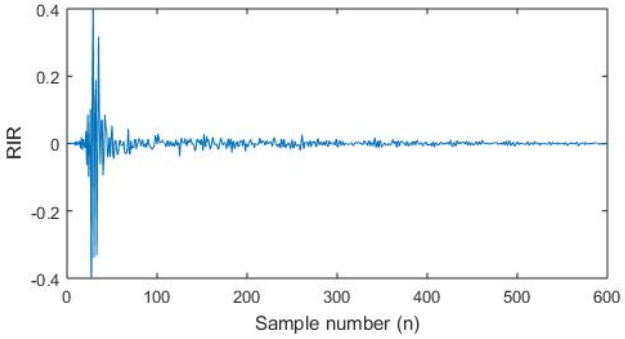

Fig. 1. A zoom-in plot of one of the RIRs from [30].

Similarly, for the single-channel speech enhancement system (WF+IMCRA), the same 15-second speech signal, as previously prepared for the multi-channel case at the same sampling frequency of 8 kHz, was convolved with an RIR obtained from one channel of the HRIR database [30]. Only one-channel of the recorded babble noise in the cafeteria environment was added to generate the corrupted microphone signal in the observed single-channel system.

For both single-channel and multi-channel systems, the noisy speech signal(s) was then analyzed using STFT with frame length of 16 ms (128 samples) and 50 % overlap, using a Hamming window. The bias compensation factor of $\beta_1 = 1.66$ and $\beta_2 = 1.0$ were selected. The length of the window function in eq.(12) was chosen as $N_b = 1$, whereas the length of another window function in eq.(14) was $N_w = 120$, with the choice of $U = 8$ and $V = 15$. The thresholds of $\gamma_0 = 4.6$ and $\zeta_0 = 1.67$, were chosen as suggested in [6]. The forgetting factors were $\alpha_v = 0.9$, $\alpha_s = 0.92$, and $\alpha = 0.9$ for tradeoff between noise reduction and speech distortion of the enhanced speech signal. The tradeoff parameter in eq.(23) was chosen as $\mu = 25$.

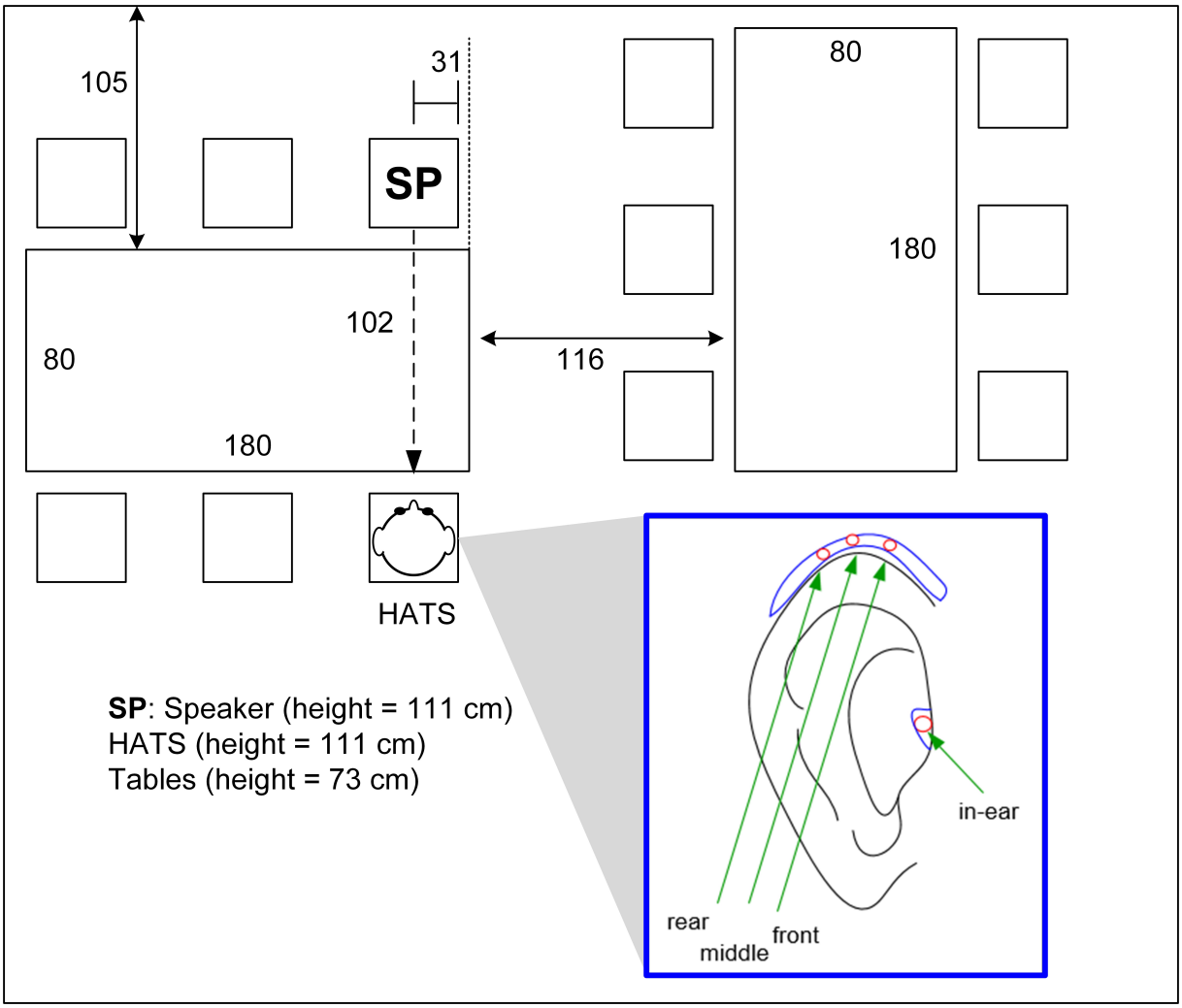

Fig. 2. A sketch of the cafeteria environment in [30]. HATS denotes the head-and-torso simulator.

For both cases, the babble noise at each microphone signal was scaled for -5, 0, and 5 dB input SNR levels.

5.2. Performance Evaluation

In order to evaluate the noise estimation performance of the investigated algorithms, the segmental noise estimation error, SegErr, defined as the squared difference between true and estimated noise covariance matrices, is employed, i.e.,

$$Seg\text{Err} = \frac{1}{L}\sum_{l=0}^{L-1}\left(\frac{\sum_k \left(\text{tr}[\hat{\mathbf{R}}_{\mathbf{vv}}(k,l) - \mathbf{R}_{\mathbf{vv}}(k,l)]\right)^2}{\sum_k \left(\text{tr}[\mathbf{R}_{\mathbf{vv}}(k,l)]\right)^2}\right). \quad (28)$$

Furthermore, plots of the true and estimated noise power spectra are also shown to evaluate the noise tracking performance.

Waveform and spectrogram plots are also included to demonstrate the overall performance of speech enhancement algorithm employing the investigated noise estimation algorithms, i.e., noise reduction performance and preservation of speech spectral components.

In addition, the SNR improvement of the enhanced speech signal was used to indicate the noise reduction performance of the speech enhancement algorithm. For a given input SNR level, the output SNR was defined by

$$\text{SNR}_{o/p} = 10 \times \log_{10}\left(\frac{\sum_{n=0}^{N-1}\hat{s}^2(n)}{\sum_{n=0}^{N-1}(s(n)-\hat{s}(n))^2}\right) \quad (29)$$

where $\hat{s}(n)$ was the enhanced speech signal and N was the total number of speech samples. The SNR improvement, ΔSNR, was the difference between the output and input SNRs of the enhanced speech signal. The higher the value of ΔSNR becomes, the better noise reduction performance it achieves. For both single-channel and multi-channel cases, the values of ΔSNR were calculated by using the information from channel 1.

Furthermore, an aspect of speech intelligibility of the enhanced signals from those investigated algorithms is observed. A short-time objective intelligibility (STOI) is a correlation-based measure between the clean and the enhanced speech signals. A higher STOI score indicates better preservation of the original speech frequency components, which means that a slighter amount of speech distortion is presented [31].

5.3. Noise Estimation Performance via SegErr

From Table 2, it is shown that SegErr of the MC-IMCRA algorithm is significantly smaller than those obtained by employing the IMCRA method; i.e., the true and estimated noises obtained by the MC-IMCRA algorithm are similar to each other. This illustrates superior noise estimation performance of the proposed multi-channel noise estimation algorithm to the single-channel one.

Table 2. Noise estimation performance of the IMCRA and MC-IMCRA techniques, via the SegErr. (babble noise)

Input SNR (dB)	SegErr (dB)	
	IMCRA	MC-IMCRA
-5	0.53	0.33
0	0.54	0.36
5	0.53	0.39

5.4. Noise Tracking Performance

To demonstrate the noise tracking performance of the proposed MC-IMCRA noise estimation algorithm, as compared to the single-channel IMCRA one, the estimated noise power spectrum was compared with its true noise spectrum. The smoothed noisy speech spectrum (red lines) was also included to identify the dominance of speech and noise. By considering at different frequency bins ($k = 40, 80, 120$ which corresponded to frequencies $0.625, 1.25$, and 1.875 kHz, respectively), when the additive noise was the babble noise at 5-dB input SNR, it was demonstrated in Fig. 3 on the left-hand-side that the MC-IMCRA algorithm was able to track the true noise power spectrum more accurately than the single-channel IMCRA one. During the speech presence (when the red lines showed high peaks), it is obvious that the noise tracking of both noise estimation algorithms was disabled; i.e., the noise estimators did not update. On the other hand, both algorithms showed rapid tracking performance of the noise spectrum when the speech was absent. However, there were some intervals that the single-channel IMCRA method cannot track the true noise spectrum. At some frames, the estimated noise power spectrum using the single-channel IMCRA algorithm was much lower than that employing the proposed MC-IMCRA one, such as during the frame indices from 280 to 440 and frome 750 to 950 for the case of $k = 40$, etc. This guarantees that the estimated noise spectrum obtained by the proposed MC-IMCRA algorithm yields better noise tracking and more accuracy than the single-channel IMCRA technique. As for the *cross*-noise power spectra between channel 1 and channel 2, its estimate by the MC-IMCRA was also illustrated in Fig. 3(a), (b), (c), on the right-hand-side. Note that, there is no *cross*-noise power spectrum for the IMCRA method.

5.5. Waveform and Spectrogram Plots

In order to investigate the effectiveness of the MWF+MC-IMCRA for speech enhancement over the WF+IMCRA, waveform and spectrogram plots of the enhanced speech signals are given in this subsection. The spectrogram plots demonstrate both the noise reduction performance and the ability to maintain the speech spectral components of the investigated speech enhancement

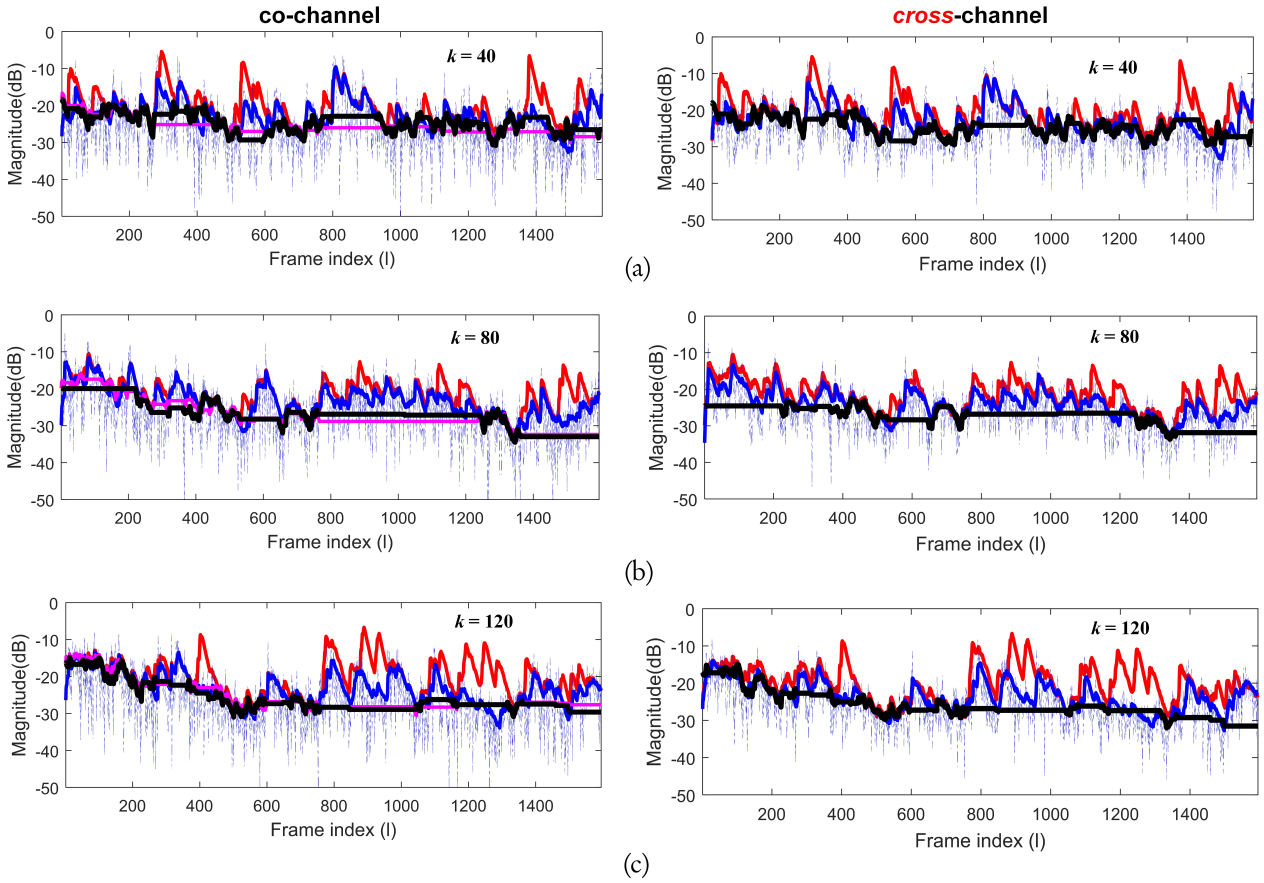

Fig. 3. Plots of smoothed noisy speech spectrum (red), true/smoothed (dotted/solid blue - left) co-noise spectrum in channel 1, true/smoothed (dotted/solid blue - right) *cross*-noise spectrum between channel 1 and 2, the estimated noise spectrum by the single-channel IMCRA method (margenta), and the estimated co/*cross*-noise spectrum by the MC-IMCRA method (black) for different frequency bins; (a) $k = 40$ (0.625 kHz), (b) $k = 80$ (1.25 kHz), and (c) $k = 120$ (1.875 kHz). (babble noise, input SNR= 5 dB) Note that, there is no *cross*-noise spectrum for the IMCRA method.

algorithm. By observing at the spectrogram plot of the enhanced speech signal using the WF+IMCRA algotithm, as illustrated in Fig. 4(c), it is evident that the ambient noise is diminished significantly. However, the low-frequency speech spectral components are practically removed. Moreover, residual noises are clearly seen particularly in the high-frequency region. Hence, this identifies that the IMCRA method overestimates the noise spectrum so that it does not only remove the noise spectral components but also some speech spectral components. On the other hand, it is clearly seen from the spectrogram plot of the enhanced speech signal using the MWF+MC-IMCRA algotithm in Fig. 4(d) that the speech spectral components are better preserved at low-frequencies. This guarantess that the noise estimate of the proposed MC-IMCRA algorithm is more accurate than its single version. Note that, it is rather difficult to remove the babble noise completely due to its speech-like and non-stationary characteristics.

5.6. Noise Reduction Performance via ΔSNR

It is shown in Table 3 that ΔSNR of the MWF+MC-IMCRA algorithm is higher than those obtained by the WF+IMCRA method for all three cases of input SNR levels. This demonstrates that the noise reduction performance of the MWF+MC-IMCRA algorithm is much better than the WF+IMCRA one. As the input SNR level is decreased, higher levels of SNR improvement is obtained. This is because the amount of additive noise is more significant, especially at low input SNR levels.

Table 3. Noise reduction performance of the IMCRA and MC-IMCRA techniques, via the SNR improvement. (babble noise)

Input SNR (dB)	ΔSNR (dB)	
	IMCRA	MC-IMCRA
-5	6.79	6.94
0	3.61	4.55
5	0.08	1.50

5.7. Speech Preservation Ability via STOI Measure

From Table 4, it is clearly shown that the STOI measures between the clean and the enhanced speech signals using the MWF+MC-IMCRA algorithm were higher than those obtained by using the WF+IMCRA method. This indicates the ability to preserve speech frequency components in the enhanced speech signals of the multi-channel speech enhancement using the proposed MC-IMCRA technique.

Table 4. Speech preservation ability of the IMCRA and MC-IMCRA techniques, via STOI measure. (babble noise)

Input SNR (dB)	STOI	
	IMCRA	MC-IMCRA
-5	0.3577	0.4151
0	0.4896	0.5467
5	0.6117	0.6645

6. Conclusions

The multi-channel extension of the IMCRA noise estimation technique has been formulated in this article. With the spatial information of microphone array signals, the proposed noise estimator is guaranteed to achieve a better noise spectral estimate, as compared to its single-channel version. Simulation results with room reverberation under a cafeteria environment and highly non-stationary babble noise have indicated that the multi-channel speech enhancement algorithm utilising the proposed multi-channel noise estimation technique outperforms that using the single-channel counterpart in terms of the noise tracking performance, noise estimation error, noise reduction performance, and the short-time objective intelligibility of the enhanced speech signal.

7. Acknowledgement

This research work is financially supported by Chula Engineering's promoting research grant, Faculty of Engineering, Chulalongkorn University.

References

[1] S. Boll, Suppression of acoustic noise in speech using spectral subtraction, *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, April 1979.

[2] B. L. Sim, Y. C. Tong, J. S. Chang and C. T. Tan, A parametric formulation of the generalized spectral subtraction method, *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 4, pp. 328–337, July 1998.

[3] H. Gustafsson, S. E. Nordholm and I. Claesson, Spectral subtraction using reduced delay convolution and adaptive averaging, *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 8, pp. 799–807, November 2001.

[4] R. Martin, Noise power spectral density estimation based on optimal smoothing and minimum statistics, *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, July 2001.

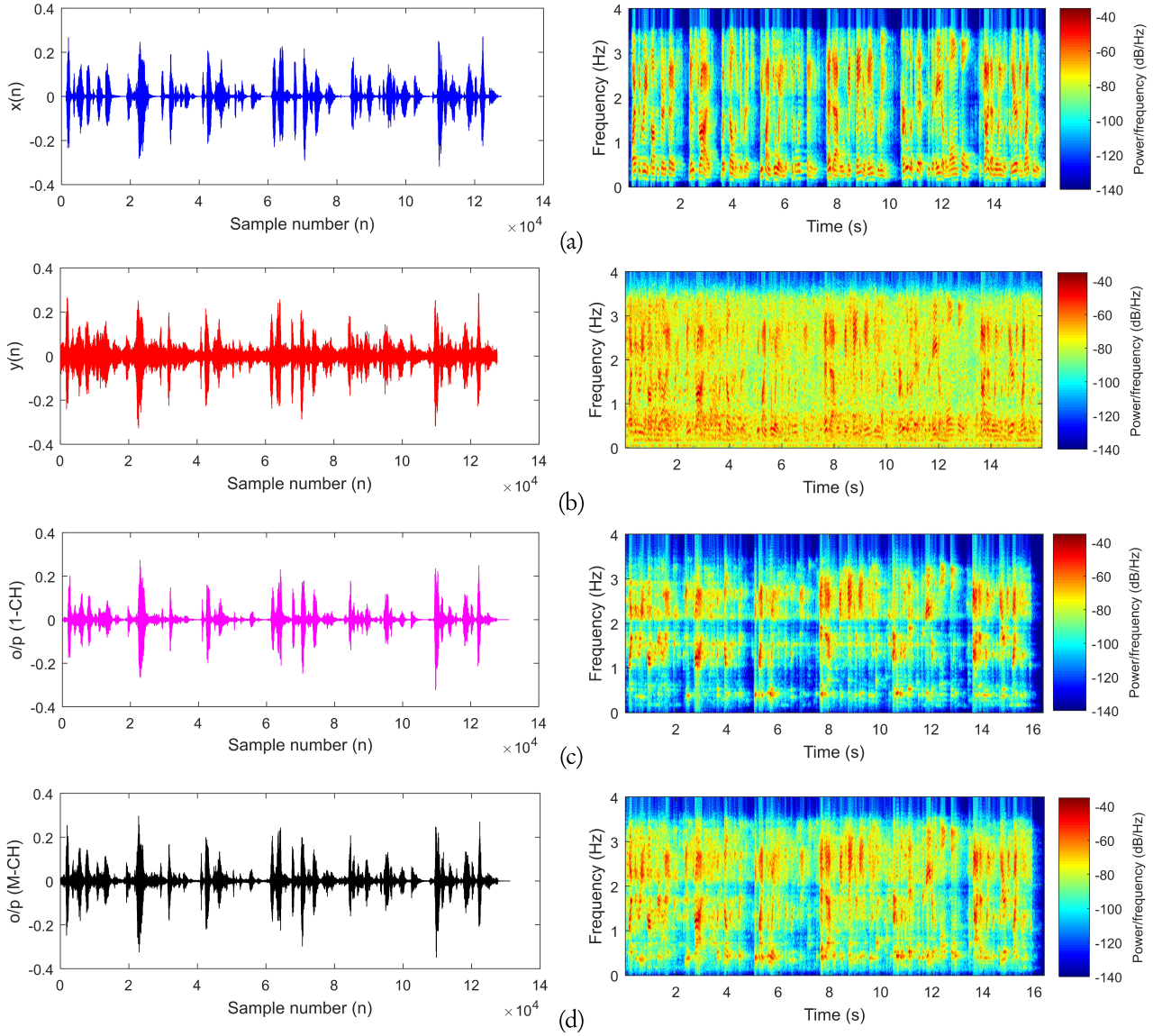

Fig. 4. Waveform and spectrogram plots of (a) the reverberant speech signal, (b) noisy and reverberant speech signal in channel 1 (babble noise, input SNR = 5 dB), (c) the enhanced speech signal using the WF+IMCRA algorithm and (d) the enhanced speech signal using the MWF+MC-IMCRA algorithm.

[5] I. Cohen and B. Berdugo, Noise estimation by minima controlled recursive averaging for robust speech enhancement, *IEEE Signal Processing Letters*, vol. 9, no. 1, pp. 12-015, January 2002.

[6] I. Cohen, Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging, *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, September 2003.

[7] N. W. D. Evans and J. S. Mason, Time-frequency quantile-based noise, in *Proc. 11th European Signal Processing Conference* (EUSIPCO), Toulouse, France, 2002, pp. 1–4.

[8] V. -K. Mai, D. Pastor, A. Aïssa-El-Bey and R. Le-Bidan, Robust estimation of non-stationary noise power spectrum for speech enhancement, *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 670–682, April 2015.

[9] N. Tiwari and P. C. Pandey, Speech enhancement using noise estimation with dynamic quantile tracking, *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2301–2312, December 2019.

[10] S. Nordholm, I. Claesson and B. Bengtsson, Adaptive array noise suppression of handsfree speaker input in cars, *IEEE Trans. on Vehicular Technology*, vol. 42, no. 4, pp. 514–518, November 1993.

[11] O. Hoshuyama, A. Sugiyama and A. Hirano, A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters, *IEEE Trans. on Signal Processing*, vol. 47, no. 10, pp. 2677–2684, October 1999.

[12] S. Markovich, S. Gannot and I. Cohen, Multichannel eigenspace beamforming in a reverberant noisy environment with multiple Interfering Speech Signals, *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1071–1086, August 2009.

[13] E. A. P. Habets, J. Benesty, S. Gannot, P. A. Naylor and I. Cohen, On the application of the LCMV beamformer to speech enhancement, in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (WASPAA), 2009, pp. 141–144.

[14] J. Benesty, J. Chen, Y. Huang and J. Dmochowski, On microphone-array beamforming from a MIMO acoustic signal processing perspective, *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1053–1065, March 2007.

[15] C. Pan, J. Chen and J. Benesty, Microphone array beamforming with high flexible interference attenuation and noise reduction, *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 30, pp. 1865–1876, May 2022.

[16] L. Griffiths and C. Jim, An alternative approach to linearly constrained adaptive beamforming, *IEEE Trans. on Antennas and Propagation*, vol. 30, no. 1, pp. 27–34, January 1982.

[17] M. Souden, J. Benesty and S. Affes, On Optimal Frequency-Domain Multichannel Linear Filtering for Noise Reduction, *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 260-276, February 2010.

[18] S. Doclo and M. Moonen, GSVD-based optimal filtering for single and multi-microphone speech enhancement, *IEEE Trans. on Signal Processing*, vol. 50, no. 9, pp. 2230–2244, September 2002.

[19] S. Doclo, A. Spriet and M. Moonen, Efficient frequency-domain implementation of speech distortion weighted multi-channel wiener filtering for noise reduction, in *Proc. 12th European Signal Processing Conference* (EUSIPCO), Vienna, Austria, 2004, pp. 2007–2010.

[20] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, Speech distortion weighted multichannel Wiener filtering techniques for noise reduction, *Speech enhancement*, Springer, pp. 199–228, 2005.

[21] B. Cornelis, M. Moonen and J. Wouters, Comparison of frequency domain noise reduction strategies based on multichannel Wiener filtering and spatial prediction, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), Taipei, Taiwan, 2009, pp. 129–132.

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

[23] R. V. Rompaey and M. Moonen, GEVD Based Speech and Noise Correlation Matrix Estimation for Multichannel Wiener Filter Based Noise Reduction, in *Proc.* 26^{th} *European Signal Processing Conference* (EUSIPCO), Rome, Italy, 2018, pp. 2544-2548.

[24] S. Bageiri and D. Giacobello, Exploiting multichannel speech presence probablity in parametric multi-channel Wiener filter, in *Proc. Interspeech Conference*, Graz, Austria, 2019, pp. 101-105.

[25] I. Cohen, Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator, *IEEE Signal Processing Letters*, vol. 9, no. 4, pp. 113–116, April 2002.

[26] P. C. Loizou, Speech Enhancement: Theory and Practice, 2nd ed., CRC Press, pp. 465–584, 2007.

[27] Y. Ephraim and D. Malah, Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator, *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, December 1984.

[28] P. Scalart and J. V. Filho, Speech enhancement based on a priori signal to noise estimation, in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing* (ICASSP), Atlanta, GA, USA, vol. 2, 1996, pp. 629–632.

[29] IEEE Subcommittee, IEEE Recommended Practice for Speech Quality Measurements, *IEEE Trans. Audio and Electroacoustics*, **AU-17(3)**, pp. 225–246, 1969.

[30] H. Kayser, S.D. Ewert, J. Anemuller and T. Rohdenburg, Database for multichannel in-ear and behind-the-ear head-related and binaural room impulse responses, *EURASIP Journal on Advances in Signal Processing*, vol. 1, no. 6, December 2009.

[31] C. H. Taal, R. C. Hendriks, R. Heusdens and J. Jensen, An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech, IEEE Trans. on Audio, Speech, and Language Processing, vol. 19, no. 7, pp. 2125–2136, September 2011.

DOI:10.4186/ej.2023.27.11.99

Nisachon Tangsangiumvisai was born in Bangkok, Thailand in 1974. She received the M.Eng. degree in Electrical and Electronic Engineering and PhD. degree in Signal Processing from the Department of Electrical and Electronics Engineering, Imperial College, London, U.K. in 1997 and 2001, respectively.

She was a recipient of the Royal Thai Scholarship to study in the U.K. during 1992-2001. After graduation, she has been with the Department of Electrical Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok, Thailand. In 2005, she obtained a research fellowship from Japan Society for the Promotion of Science (JSPS) to conduct her research at Tokyo Institute of Technology, Japan. In 2011, she has become an Associate Professor with the Department of Electrical Engineering, Faculty of Engineering, Chulalongkorn University. Her research interests include Adaptive Signal Processing, Noise Reduction techniques for Speech Enhancement, etc.