Zou et al. EURASIP Journal on Audio, Speech, and Music Processing 2014, 2014:17
http://asmp.eurasipjournals.com/content/2014/1/17

EURASIP Journal on Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Speech enhancement with an acoustic vector sensor: an effective adaptive beamforming and post-filtering approach

Yue Xian Zou[1*], Peng Wang[1], Yong Qing Wang[1], Christian H Ritz[2] and Jiangtao Xi[2]

Abstract

Speech enhancement has an increasing demand in mobile communications and faces a great challenge in a real ambient noisy environment. This paper develops an effective spatial-frequency domain speech enhancement method with a single acoustic vector sensor (AVS) in conjunction with minimum variance distortionless response (MVDR) spatial filtering and Wiener post-filtering (WPF) techniques. In remote speech applications, the MVDR spatial filtering is effective in suppressing the strong spatial interferences and the Wiener post-filtering is considered as a popular and powerful estimator to further suppress the residual noise if the power spectral density (PSD) of target speech can be estimated properly. With the favorable directional response of the AVS together with the trigonometric relations of the steering vectors, the closed-form estimation of the signal PSDs is derived and the frequency response of the optimal Wiener post-filter is determined accordingly. Extensive computer simulations and a real experiment in an anechoic chamber condition have been carried out to evaluate the performance of the proposed algorithm. Simulation results show that the proposed method offers good ability to suppress the spatial interference while maintaining comparable log spectral deviation and perceptual evaluation of speech quality performance compared with the conventional methods with several objective measures. Moreover, a single AVS solution is particularly attractive for hands-free speech applications due to its compact size.

Keywords: Speech enhancement; Acoustic vector sensor; Beamforming; Post-filtering; Power spectral density estimation

1 Introduction

As the presence of background noise significantly deteriorates the quality and intelligibility of speech, enhancement of speech signals has been an important and challenging problem and various methods have been proposed in the literature to tackle this problem. Spectral subtraction, Wiener filtering, and their variations [1] are commonly used for suppressing additive noise, but they are not able to effectively suppress spatial interference. In order to eliminate spatial interferences, beamforming techniques applied to microphone array recordings can be employed [2-9]. Among these, the minimum variance distortionless response (MVDR) beamformer known as the Capon beamformer and their equivalent generalized sidelobe cancellers (GSC) work successfully in remote speech enhancement applications [2]. However, the performance of MVDR-type methods is proportional to the number of array sensors used, thus limiting their application. Moreover, the MVDR beamformer is not effective at suppressing additive noise, leaving residual noise in its output. As a result, the well-known Wiener post-filtering solution normally can be employed to further reduce the residual noise from the output of the beamformer [7]. Recently, speech enhancement using the acoustic vector sensor (AVS) array has received research attention due to the merit of spatial co-location of microphones and signal time alignment [5,10-12]. Compared with the traditional microphone array, the compact structure (occupying a volume of approximately 1 cm^3) makes the AVS much more attractive in portable speech enhancement applications. Research showed that the AVS array beamformer with the MVDR method [5,10] successfully suppresses spatial interferences but fails to effectively suppress background noise. The integrated MVDR and Wiener post-filtering method

* Correspondence: zouyx@pkusz.edu.cn
[1]ADSPLAB/ELIP, School of Electronic Computer Engineering, Peking University, Shenzhen 518055, China
Full list of author information is available at the end of the article

using AVS array [12] offers good performance in terms of suppression of spatial interferences and background additive noise, but it requires more than two AVS units as well as the good voice activity detection (VAD) technique.

In this paper, we focus on developing a speech enhancement solution capable of effectively suppressing spatial interferences and additive noise at a less computational cost using only one AVS unit. More specifically, by exploring the unique spatial co-location property (the signal arrives at the sensors at the same time) and the trigonometric relations of the steering vectors of the AVS, a single AVS-based speech enhancement system is proposed. The norm-constrained MVDR method is employed to form the spatial filter, while the optimal Wiener post-filter is designed by using a novel closed-form power spectral density (PSD) estimation method. The proposed solution does not depend on the VAD technique (for noise estimation) and hence has advantages of small size, less computation cost, and the ability to suppress both spatial interferences and background noise.

The paper is organized as follows. The data model of an AVS and the frequency domain MVDR (FMV) with a single AVS are presented in Section 2. The detailed derivation of the closed-form estimation of the signal PSDs for an optimal Wiener post-filtering (WPF) using the AVS structure is given in Section 3. The proposed norm-constrained FMV-effective Wiener post-filtering (NCFMV-EWPF) algorithm for speech enhancement is presented in Section 4. Simulation results are presented in Section 5. Section 6 concludes our work.

2 Problem formulation

2.1 Data model for an AVS unit

An AVS unit generally consists of four co-located constituent sensors, including one omnidirectional sensor (denoted as the *o*-sensor) and three orthogonally oriented directional sensors depicted as the *u*-sensor, *v*-sensor, and *w*-sensor, respectively. As an example, Figure 1 shows a data capture system with an AVS unit. In this paper, focusing on deriving the algorithm and making the derivation clear, let us assume that there is one target speech $s(t)$ at $(\theta_s,\ \phi_s) = (90°,\ \phi_s)$ and one interference signal $s_i(t)$ at $(\theta_i,\ \phi_i) = (90°,\ \phi_i)$ impinging on this AVS unit, where $\phi_s,\ \phi_i \in [0°,\ 360°)$ are the azimuth angles. Since $s(t)$ and $s_i(t)$ arrive in the horizontal plane, we only need the *u*-sensor, *v*-sensor, and *o*-sensor to capture signals from the AVS unit. The angle difference between $s(t)$ and $s_i(t)$ is defined as

$$\Delta\phi = \phi_s - \phi_i \tag{1}$$

The corresponding steering vectors are given by

$$\boldsymbol{v}(\phi_s) = [u_s, v_s, 1]^T = [\cos\phi_s,\ \sin\phi_s, 1]^T \tag{2}$$

$$\boldsymbol{v}(\phi_i) = [u_i, v_i, 1]^T = [\cos\phi_i,\ \sin\phi_i, 1]^T \tag{3}$$

where $[.]^T$ denotes the vector/matrix transposition.

In the cases that room reverberation does not exist, the received data of the AVS can be modeled as [13]

$$\boldsymbol{x}_{\text{avs}}(t) = \boldsymbol{v}(\phi_s)s(t) + \boldsymbol{v}(\phi_i)s_i(t) + \boldsymbol{n}_{\text{avs}}(t) \tag{4}$$

where $\mathbf{n}_{\text{avs}}(t)$ is assumed as the additive white Gaussian noise at the AVS unit. Specifically, we have the following definitions:

$$\boldsymbol{x}_{\text{avs}}(t) = [x_u(t), x_v(t), x_o(t)]^T \tag{5}$$

$$\boldsymbol{n}_{\text{avs}}(t) = [n_u(t), n_v(t), n_o(t)]^T \tag{6}$$

where $x_u(t)$, $x_v(t)$, and $x_o(t)$ are the received data of the *u*-, *v*-, and *o*-sensor, respectively, and $n_u(t)$, $n_v(t)$, and $n_o(t)$ are the captured noise at the *u*-, *v*-, and *o*-sensor, respectively. The task of speech enhancement with an AVS is to estimate $s(t)$ from $\boldsymbol{x}_{avs}(t)$.

In this study, without loss of generality, we follow the commonly used assumptions [4]: (1) $s(t)$ and $s_i(t)$ are mutually uncorrelated; (2) $n_u(t)$, $n_v(t)$, and $n_o(t)$ are mutually uncorrelated.

2.2 FMV with a single AVS

The MVDR beamformer is considered as one of the most popular beamforming methods for suppressing spatial interferences in remote speech applications. In this subsection, we present the formulation of the frequency domain MVDR beamformer (FMV) with two sensors (*u*-sensor and *v*-sensor) of the AVS unit. From (2) to (4), the data received by the *u*-sensor and the *v*-sensor can be modeled as [14]

$$\boldsymbol{x}(t) = [x_u(t), x_v(t)]^T = \boldsymbol{a}(\phi_s)s(t) + \boldsymbol{a}(\phi_i)s_i(t) + \boldsymbol{n}(t) \tag{7}$$

where

$$\boldsymbol{a}(\phi_s) = [u_s, v_s]^T = [\cos\phi_s,\ \sin\phi_s]^T \tag{8}$$

and

$$\boldsymbol{a}(\phi_i) = [u_i, v_i]^T = [\cos\phi_i,\ \sin\phi_i]^T \tag{9}$$

The frequency domain formulation of the data model of (7) is given by

$$\boldsymbol{X}(f) = \boldsymbol{a}(\phi_s)S(f) + \boldsymbol{a}(\phi_i)S_i(f) + \boldsymbol{N}(f) \tag{10}$$

where $\boldsymbol{X}(f) = [X_u(f), X_v(f)]^T$ and $\boldsymbol{N}(f) = [N_u(f), N_v(f)]^T$. The beamforming is then performed by applying a complex

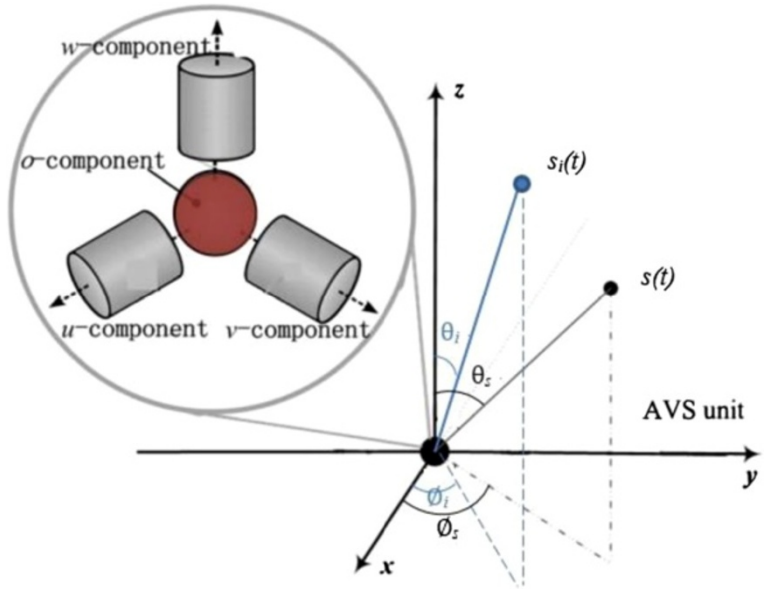

Figure 1 Illustration of one AVS data capturing system [13,14].

weight to the captured signals, and the output of the FMV can be denoted as

$$Y(f) = \boldsymbol{w}^H(f)\boldsymbol{X}(f) = \boldsymbol{w}^H(f)\boldsymbol{a}(\phi_s)S(f) + \boldsymbol{w}^H(f)(\boldsymbol{a}(\phi_i)S_i(f) + \boldsymbol{N}(f)) \tag{11}$$

where $(.)^H$ denotes the Hermitian transposition. $\boldsymbol{w}^H(f) = [w_u(f), w_v(f)]$ is the weight vector of the FMV. Let us define

$$g(\phi_s,f) = \boldsymbol{w}^H(f)\boldsymbol{a}(\phi_s) \tag{12}$$

$$g(\phi_i,f) = \boldsymbol{w}^H(f)\boldsymbol{a}(\phi_i) \tag{13}$$

Obviously, $g(\phi_s,f)$ and $g(\phi_i,f)$ can be viewed as the spatial response gains of the FMV to the target spatial signal $S(f)$ and the spatial interference signal $S_i(f)$, respectively. Substituting (12) and (13) into (11), we can get

$$Y(f) = \boldsymbol{w}^H(f)\boldsymbol{X}(f) = g(\phi_s,f)S(f) + g(\phi_i,f)S_i(f) + \boldsymbol{w}^H(f)\boldsymbol{N}(f) \tag{14}$$

The basic idea of designing the optimal FMV is to maintain $g(\phi_s,f) = 1$ for $S(f)$ while minimizing the output signal power ($P_{YY} = E[Y(f)Y^*(f)]$) of the FMV to suppress other undesired sources. Hence, the optimal weight vector of the FMV can be obtained by solving the constrained optimization cost function [2]:

$$\boldsymbol{w}_{\mathrm{FMV}}(f) = \underset{\boldsymbol{w}}{\arg\min}\ P_{YY} \quad \text{subject to } g(\phi_s,f) = 1, \text{and } P_{YY} = \boldsymbol{w}^H(f)\boldsymbol{R}_{\boldsymbol{x}}(f)\boldsymbol{w}(f) \tag{15}$$

where $\mathbf{R}_{\boldsymbol{x}}(f) = E[\boldsymbol{X}(f)\boldsymbol{X}^H(f)]$ is the autocorrelation matrix of the received data of the FMV. The optimal solution of (15) is given as [2]

$$\boldsymbol{w}_{\mathrm{FMV}}(f) = \frac{\boldsymbol{R}_{\boldsymbol{x}}^{-1}(f)\boldsymbol{a}(\phi_s)}{\boldsymbol{a}^T(\phi_s)\boldsymbol{R}_{\boldsymbol{x}}^{-1}(f)\boldsymbol{a}(\phi_s)} \tag{16}$$

Equation 16 is the standard form of the FMV. It is clear that when $\boldsymbol{a}(\phi_s)$ is fixed (speech target is static), $\boldsymbol{w}_{\mathrm{FMV}}(f)$ depends on the estimate of $\boldsymbol{R}_{\boldsymbol{x}}^{-1}(f)$. There are several methods that have been proposed to estimate $\boldsymbol{R}_{\boldsymbol{x}}(f)$ [1], and the diagonal loading technique is one of the robust algorithms aiming at avoiding the non-singularity in (16), which leads to a norm-constrained FMV (NCFMV) as shown in (17) [3]:

$$\boldsymbol{w}_{NC}(f) = \frac{(\boldsymbol{R}_{\boldsymbol{x}}(f) + \gamma\boldsymbol{I})^{-1}\boldsymbol{a}(\phi_s)}{\boldsymbol{a}^T(\phi_s)(\boldsymbol{R}_{\boldsymbol{x}}(f) + \gamma\boldsymbol{I})^{-1}\boldsymbol{a}(\phi_s) + \sigma} \tag{17}$$

where $\boldsymbol{I}$ is an identity matrix, γ is the positive loading factor, and σ is a small positive number to avoid the denominator becoming zero. It is expected that the NCFMV will greatly suppress the spatial unwanted signals. Obviously, the output of the NCFMV can be derived as follows with (17), (12), (13), and some simple manipulations:

$$Y(f) = \boldsymbol{w}_{NC}^H(f)\boldsymbol{X}(f) = S(f) + \boldsymbol{w}_{NC}^H(f)\boldsymbol{a}(\phi_i)S_i(f) + \boldsymbol{w}_{NC}^H(f)\boldsymbol{N}_{un}(f) = S(f) + g(\phi_i,f)S_i(f) + \boldsymbol{w}_{NC}^H(f)\boldsymbol{N}_{un}(f) \tag{18}$$

2.3 The estimation of the power spectral density

As discussed above, the NCFMV is only effective in suppressing the spatial interferences. In this section, a new solution has been proposed by incorporating the well-known Wiener post-filter (WPF) to further suppress the residual noise in beamformer output $Y(f)$ in (18).

Basically, according to the formulation of the Wiener filter in the frequency domain, to estimate $S(f)$ from $Y(f)$, the frequency response of the Wiener filter is given by [6,8]

$$W_{pf}(f) = \psi_{YS}(f)/\psi_{YY}(f) \approx \psi_{SS}(f)/\psi_{YY}(f) \tag{19}$$

where $\psi_{YS}(f)$ is the cross-power spectrum density (CSD) of $S(f)$ and $Y(f)$ and $\psi_{YY}(f)$ is the power spectral density (PSD) of $Y(f)$. Generally, $Y(f)$ are considered uncorrelated to interferences, and we can approximately get the second equation in (19) via (18). From (19), it is clear that a good estimate of $\psi_{SS}(f)$ and $\psi_{YY}(f)$ from $X(f)$ and $Y(f)$ are very crucial to the performance of the WPF. There are some PSD estimation algorithms that have been proposed under different spatial-frequency joint estimation schemes. For single-channel application as an example, the voice activity detection (VAD) method is usually applied to get the noise and speech segments, and then the spectrum subtraction algorithm can be used to remove noise components before estimating $\psi_{SS}(f)$. Moreover, for microphone array post-filtering schemes, $\psi_{SS}(f)$ can be estimated from the available multichannel signals, which are assumed to be within an incoherent noise environment [6].

Motivated by the unique properties of the AVS, where multichannel signals are available (u-, v, and o-sensor signals) and there exists a trigonometric relationship between the steering vectors $\boldsymbol{a}(\phi_s)$ and $\boldsymbol{a}(\phi_i)$ of the AVS, in this paper, we will derive a closed-form solution to estimate $\psi_{SS}(f)$ and $\psi_{YY}(f)$ to form an optimal WPF. The system diagram proposed is shown in Figure 2.

3 The formulation of the Wiener post-filter

3.1 Derivation of the estimate of CSD and PSD

For presentation clarity, let us define the notation of the cross-power spectral density (CSD) between $\alpha(f)$ and $\beta(f)$ as

$$\psi_{\alpha\beta}(f) = E[\alpha(f)\beta^*(f)] \tag{20}$$

From (10), we have

$$X_u(f) = \cos(\phi_s)S(f) + \cos(\phi_i)S_i(f) + N_u(f) \tag{21}$$

$$X_v(f) = \sin(\phi_s)S(f) + \sin(\phi_i)S_i(f) + N_v(f) \tag{22}$$

For presentation simplicity, the frequency index f will be dropped in the following derivation. Ideally, the additive noises of u-, v-, and o-sensors have the same power, and then we have

$$\psi_{NN} = E\left[N_u N_u^*\right] = E\left[N_v N_v^*\right] = E\left[N_o N_o^*\right] \tag{23}$$

It is noted that the assumption of equal power for all channels used in (23) is not true for the real signals recorded by the AVS unit, but these can be calibrated in practice [15]. With (18), (21), (22), and (23), the CSD and PSD of signals can be derived following the definition given in (20):

$$\begin{aligned}\psi_{uu} = E\left[X_u X_u^*\right] &= \cos^2(\phi_s)E[SS^*] + \cos^2(\phi_i)E\left[S_i S_i^*\right] \\ &+ E\left[N_u N_u^*\right] = \cos^2(\phi_s)\psi_{SS} + \cos^2(\phi_i)\psi_{S_iS_i} + \psi_{NN}\end{aligned} \tag{24}$$

$$\begin{aligned}\psi_{vv} = E\left[X_v X_v^*\right] &= \sin^2(\phi_s)E[SS^*] + \sin^2(\phi_i)E\left[S_i S_i^*\right] \\ &+ E\left[N_v N_v^*\right] = \sin^2(\phi_s)\psi_{SS} + \cos^2(\phi_i)\psi_{S_iS_i} + \psi_{NN}\end{aligned} \tag{25}$$

$$\psi_{oo} = E\left[X_o X_o^*\right] = \psi_{SS} + \psi_{S_iS_i} + \psi_{NN} \tag{26}$$

$$\psi_{YY} = E[YY^*] = \psi_{SS} + g^2(\phi_i)\psi_{S_iS_i} + ||\boldsymbol{w}_{NC}||^2\psi_{NN} \tag{27}$$

$$\psi_{u+v} = \psi_{uu} + \psi_{vv} = \psi_{SS} + \psi_{S_iS_i} + 2\psi_{NN} \tag{28}$$

$$\psi_{Yo} = E\left[YX_o^*\right] = \psi_{SS} + g^*(\phi_i)\psi_{S_iS_i} \tag{29}$$

$$\psi_{oY} = E[X_o Y^*] = \psi_{SS} + g(\phi_i)\psi_{S_iS_i} \tag{30}$$

From (24) to (30), it is clear that there are seven equations with four unknown variables ψ_{NN}, $g(\phi_i)$, $\psi_{S_iS_i}$, and $\psi_{S_iS_i}$. Hence, using (28) and (26), the PSD of noise can be derived as

$$\psi_{NN} = \psi_{u+v} - \psi_{oo} \tag{31}$$

Similarly, the gain response of the NCFMV on the interference S_i can be given by

$$g(\phi_i) = (\psi_{oY} - \psi_{YY} + ||\boldsymbol{w}_{NC}||^2\psi_{NN})/(\psi_{oo} - \psi_{Yo} - \psi_{NN}) \tag{32}$$

Moreover, the PSD of the interference S_i and the target speech S can be derived, respectively, as follows:

$$\psi_{S_iS_i} = (\psi_{oo} - \psi_{oY} - \psi_{NN})/(1 - g(\phi_i)) \tag{33}$$

$$\psi_{SS} = \psi_{oY} - g(\phi_i)\psi_{S_iS_i} \tag{34}$$

3.2 The proposed EWPF method and some discussions

Till now, we have mathematically derived the closed-form expressions of the ψ_{SS} in (34), ψ_{YY} in (27), and W_{pf} in (19). Since Y, X_u, X_v, and X_o can be measured, the estimates of ψ_{SS} and ψ_{YY} can be determined accordingly. Hence, (33), (34), (27), and (19) describe the basic form of

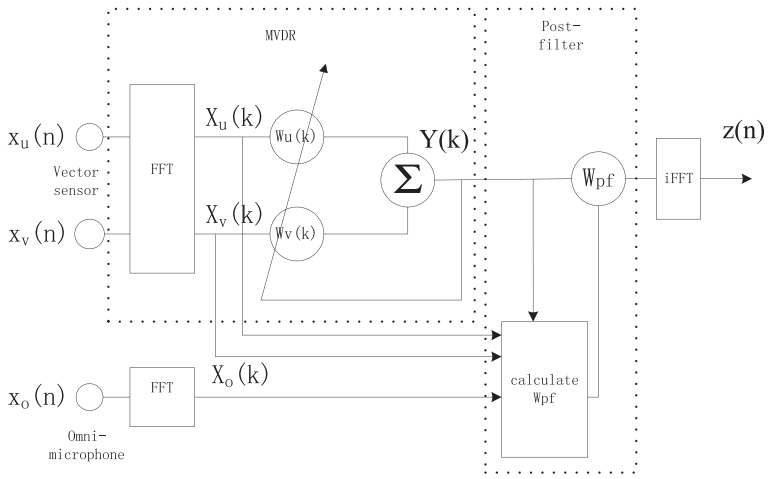

Figure 2 Our proposed speech enhancement system using single AVS.

our proposed effective Wiener post-filtering algorithm for further enhancing the speech with an AVS (here, we term it as EWPF for short). In the following context, we will have some discussions on our proposed EWPF method.

Firstly, to implement the EWPF, the cross-power spectral density $\psi_{\alpha\beta}(f)$ needs to be estimated. It is well known that the recursive update formula is a popular approach:

$$\hat{\psi}_{\alpha\beta}(f,l) = \lambda\hat{\psi}_{\alpha\beta}(f,l-1) + (1-\lambda)\alpha(f,l)\beta^{*}(f,l) \qquad (35)$$

where l is the frame index and $\lambda \in (0, 1]$ is the forgetting factor.

Secondly, it is noted that when $\Delta\phi$ defined in (1) is close to or equal to 0, the denominator in (32) goes to 0. To avoid this situation, one small positive factor σ_r should be added to the denominator of (32) and we get

$$\hat{g}(\phi_i) = \left(\hat{\psi}_{oY} - \hat{\psi}_{YY} + \|\boldsymbol{w}_{NC}\|^2\hat{\psi}_{NN}\right)/\left(\hat{\psi}_{oo} - \hat{\psi}_{Yo} - \hat{\psi}_{NN} + \sigma_r\right) \qquad (36)$$

Thirdly, analyzing the properties of $g(\phi_i)$, we observe the following: (1) If the target source $s(t)$ is considered as short-time spatially stationary (approximately true for speech applications), $\boldsymbol{w}_{NC}$ in (17) can be updated every L_u frames for reducing computational complexity. Therefore, from the definition of (13), the gain $g(\phi_i)$ will remain unchanged within L_u frames. However, $\hat{\psi}_{\alpha\beta}(f,l)$ is estimated frame by frame via (35); therefore, a more accurate estimation of $g(\phi_i)$ can be achieved by averaging over L_u frames. (2) From (36), it is clear that the small denominator will lead to a large variation of $g(\phi_i)$, reflecting incorrect estimates since the NCFMV is designed to suppress rather than to amplify the interference. Hence, it is reasonable to apply a clipping function $f_c(x, b)$ (see (43)) to remove the outliers in the estimate of $\hat{g}(\phi_i)$.

4 The proposed NCFMV-EWPF algorithm

Similar to the existing remote speech enhancement applications, our proposed algorithm is implemented in the frequency domain by segmenting the received signal into frames and then the short-time Fourier transforms (STFTs) are applied. Specifically, to determine $\boldsymbol{w}_{NC}$ in (17), the estimate of the $\boldsymbol{R}_{\boldsymbol{x}}$ is given by [10]

$$\hat{\boldsymbol{R}}_{\boldsymbol{x}}(k) = \frac{1}{F}\begin{bmatrix} C_d A(k) & B(k) \\ C(k) & C_d D(k) \end{bmatrix} \qquad (37)$$

where k is the frequency bin index and $k = 1,2,\ldots, K$. C_d is a constant slightly greater than the one that helps avoid matrix singularity. F is the frame number used for estimating $\boldsymbol{R}_{\boldsymbol{x}}(k)$, and in our study it is set as $F = 2L_u$. Let us define $X_u(k,l)$ and $X_v(k,l)$ as the kth component of the spectrum of the lth frame of $x_u(n)$ and $x_v(n)$, respectively, and we have

$$A(k) = \sum_{i=0}^{F-1} X_u^{*}(k,l-i)X_u(k,l-i) \qquad (38)$$

$$B(k) = \sum_{i=0}^{F-1} X_u^{*}(k,l-i)X_v(k,l-i) \qquad (39)$$

$$C(k) = \sum_{i=0}^{F-1} X_v^{*}(k,l-i)X_u(k,l-i) \qquad (40)$$

$$D(k) = \sum_{i=0}^{F-1} X_v^{*}(k,l-i)X_v(k,l-i) \qquad (41)$$

From (37) to (41), we can see the autocorrelation matrix $\boldsymbol{R}_{\boldsymbol{x}}(k)$ is estimated by using the F most recent fast

Fourier transforms (FFTs). Therefore, the robust estimation of $W_{pf}(k)$ in (19) asks for the robust estimation of $g(\phi_i,k)$. According to the discussions in Section 3.2, we adopt the following estimation:

$$\hat{g}(\phi_i,k)=\frac{1}{L_u}\sum_{l=L_1}^{L_2} f_c\left(\frac{\hat{\psi}_{oY}(k,l)-\hat{\psi}_{YY}(k,l)+\|\mathbf{w}_{NC}(k,l)\|^2\hat{\psi}_{NN}(k,l)}{\hat{\psi}_{oo}(k,l)-\hat{\psi}_{Yo}(k,l)-\hat{\psi}_{NN}(k,l)+\sigma_r},b\right) \tag{42}$$

where $L_1=\mathrm{fix}((l-1)/L_u)L_u+1$, $L_2=\mathrm{fix}((l-1)/L_u)L_u+L_u$, fix(.) is the floor operation, b is a predefined threshold, and $f_c(x,b)$ is the clipping function and defined as

$$f_c(x,b)=x \text{ when } 0<x\le b \text{ else } f_c(x,b)=0 \tag{43}$$

For presentation completeness, the proposed NCFMV-EWPF algorithm is summarized in Algorithm 1.

5 Simulation study

The performance evaluation of our proposed NCFMV-EWPF algorithm has been carried out in this section. The commonly used performance measurement metrics have been adopted, which include the following:

1. Output signal to interference plus noise ratio (SINR) defined as [7]

$$\mathrm{SINR}=10\log\left(\|z_s(t)\|^2/\|x_o(t)-z_s(t)\|^2\right) \tag{44}$$

where $z_s(t)$ is the enhanced speech of the system and $x_o(t)$ is the received signal of the o-sensor. Moreover, a segmental output SINR is calculated on a frame-by-frame basis and then averaged over the total number frames to get more accurate prediction of perceptual speech quality [7].

2. Log spectral deviation (LSD), which is used to measure the speech distortion and defined as [16]

$$\mathrm{LSD}=\left\|\ln\left(\psi_{ss}(f)/\psi_{zz}(f)\right)\right\| \tag{45}$$

Algorithm 1 The pseudo-code of the proposed NCFMV-EWPF algorithm

Step 1: Signal segmentation: Hamming window with 50% overlapped

Step 2: Start $l = 1{:}F$:

2.1 Compute STFTs: $X_o(k,l)$, $X_v(k,l)$, $X_v(k,l)$

2.2 If mod(l, F/2):

Calculate the estimate of $\hat{\boldsymbol{R}}_x(k)$ by (37)

Compute $\boldsymbol{w}_{NC}(k)$ by (17)

End if

2.3 Compute the beamformer output: $Y(k,l)=\boldsymbol{w}_{NC}^H(k)\boldsymbol{X}(k,l)$

2.4 Using (35), compute $\hat{\psi}_{YY}(k,l)=\lambda\hat{\psi}_{YY}(k,l-1)+(1-\lambda)Y(k,l)Y^*(k,l)$

2.5 Similarly, compute $\hat{\psi}_{oY}(k,l)$, $\hat{\psi}_{Yo}(k,l)$, $\hat{\psi}_{oo}(k,l)$, $\hat{\psi}_{uu}(k,l)$, $\hat{\psi}_{vv}(k,l)$

2.6 Compute $\hat{\psi}_{u+v}(k,l)$ by (28) and $\hat{\psi}_{NN}(k,l)$ by (31)

2.7 If mod(l, L_u):

Calculate $\hat{g}(\phi_i,k)$ by (42)

End if

2.8 For $l=l-L_u+1{:}l$:

Calculate $\hat{\psi}_{SS}(k,l)$ by (34), Compute $\hat{W}_{pf}(k,l)$ by (19)

End for l

2.9 Compute the WPF filter output in frequency domain

$Z(k,l)=\hat{\boldsymbol{W}}_{pf}(k,l)Y(k,l)$

End l %start

Step 3: Compute the output in time domain:

Using the inverse FFT and performing an overlap, add the frames to generate the time domain output.

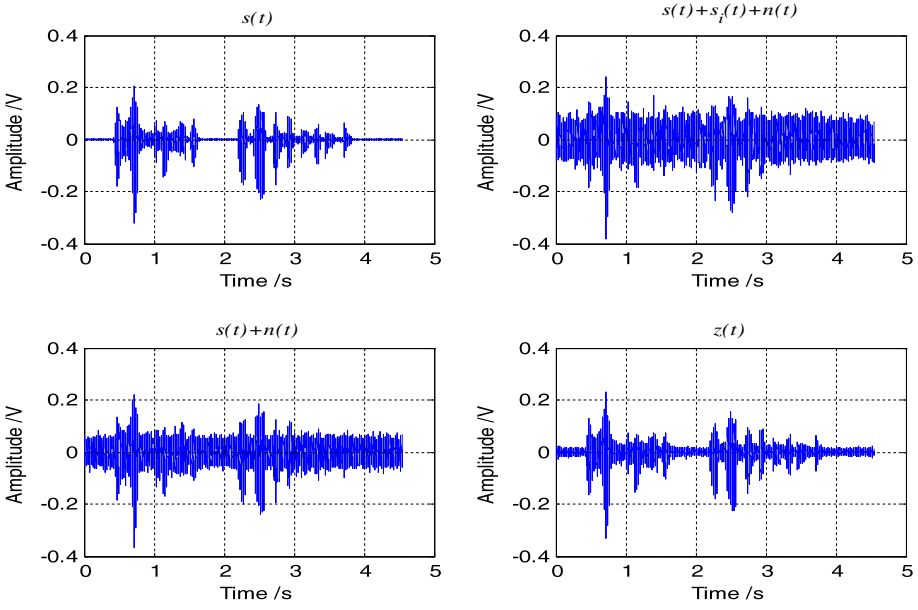

Figure 3 Waveforms. $s(t)$ is the target speech, $s_i(t)$ is the interference speech, $n(t)$ is the additive noise, and $z(t)$ is the enhanced speech signal.

where $\psi_{ss}(f)$ is the PSD of the target speech and $\psi_{zz}(f)$ is the PSD of the enhanced speech. It is clear that the smaller LSD indicates the less speech distortion. Similar to the calculation of SINR, the segmental LSD is computed.

3. Perceptual evaluation of speech quality (PESQ) [17]: To evaluate the performance of the speech enhancement algorithms, ITU-PESQ software [17] is utilized.

In addition, we also compared the performance of the Zelinski post-filter (ZPF) [4], NCFMV [5], and NCFMV-ZPF [6] algorithms under the same conditions to our proposed algorithm. The setup of the single AVS unit is shown in Figure 1.

In computer simulation studies, for each trial, a male speech lasting about 5 s acts as the target speech $s(t)$ and babble speech taken from the Noisex-92 database [18] acts as the interference speech $s_i(t)$. One set of the typical waveforms used in our simulation studies is shown in Figure 3.

5.1 Experiments on simulated data

5.1.1 Experiment 1: the SINR performance under different noise conditions

In this experiment, we have carried out nine trials (numbered as trial 1 to trial 9) to evaluate the performance of the algorithms under different spatial and additive noise conditions [9]. The experimental settings are as follows: The sampling rate is set to be 16 kHz and a 512-point FFT is used. The target speaker is located at (90°, 45°) and the interference speaker is set at (90°, 0°). For the proposed NCFMV-EWPF algorithm, parameters are set as $\lambda = 0.6$, $\sigma_r = 10^{-3}$, $L_u = 4$, $\gamma = \sigma = 10^{-5}$,

Table 1 SINR-out for different algorithms (dB)

Algorithm	ZPF [4]	NCFMV [5]	NCFMV-ZPF [6]	NCFMV-EWPF	SINR-input (dB)
Trial 1 ($\boldsymbol{n}_{avs}(t) = 0$ and $s_i(t) \neq 0$)	2.7	12.6	14.8	*26.2*	0
Trial 2 ($\boldsymbol{n}_{avs}(t) = 0$ and $s_i(t) \neq 0$)	7.8	12.8	16.4	*34.0*	5
Trial 3 ($\boldsymbol{n}_{avs}(t) = 0$ and $s_i(t) \neq 0$)	13.1	13.4	18.3	*28.3*	10
Trial 4 ($\boldsymbol{n}_{avs}(t) \neq 0$ and $s_i(t) = 0$)	8.1	2.0	7.8	*8.3*	0
Trial 5 ($\boldsymbol{n}_{avs}(t) \neq 0$ and $s_i(t) = 0$)	*13.5*	6.5	*13.5*	13.2	5
Trial 6 ($\boldsymbol{n}_{avs}(t) \neq 0$ and $s_i(t) = 0$)	*17.6*	9.1	*17.0*	16.5	10
Trial 7 ($\boldsymbol{n}_{avs}(t) \neq 0$ and $s_i(t) \neq 0$)	3.1	8.1	11.9	*14.2*	0
Trial 8 ($\boldsymbol{n}_{avs}(t) \neq 0$ and $s_i(t) \neq 0$)	8.3	10.3	14.7	*18.0*	5
Trial 9 ($\boldsymbol{n}_{avs}(t) \neq 0$ and $s_i(t) \neq 0$)	13.6	12.4	18.9	*21.2*	10

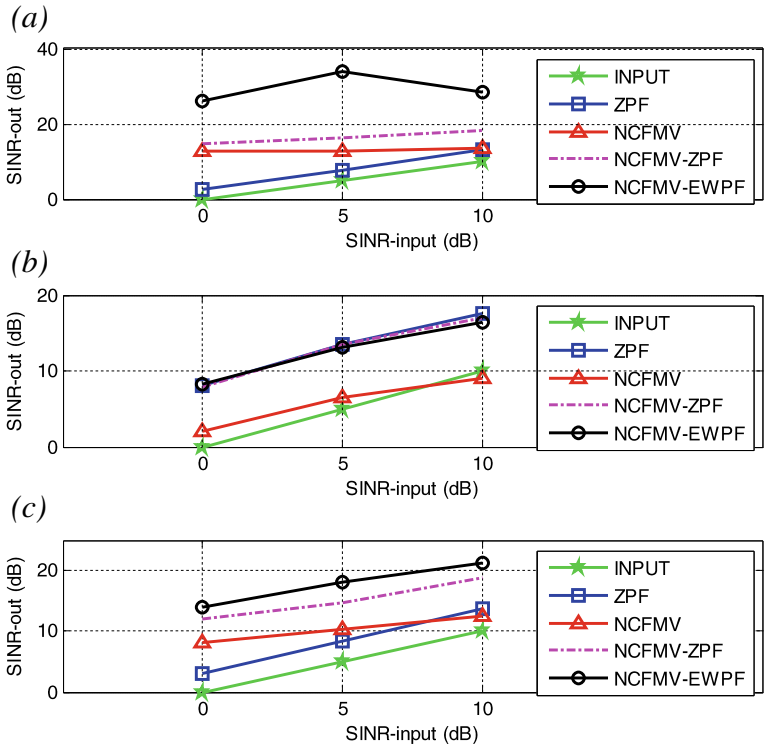

Figure 4 The performance of SINR-out versus SINR-input. (a) $\boldsymbol{n}_{avs}(t) = 0$ and $s_i(t) \neq 0$. **(b)** $\boldsymbol{n}_{avs}(t) \neq 0$ and $s_i(t) = 0$. **(c)** $\boldsymbol{n}_{avs}(t) \neq 0$ and $s_i(t) \neq 0$.

$C_d = 1.1$, and $b = 6$, which produced the best experimental results under this specific setup. For comparison algorithms, the parameter settings are set as the same as those in the relevant papers. The experimental results are listed in Table 1.

As shown in Table 1, the best performance for different conditions is addressed in italics. The proposed NCFMV-EWPF algorithm outperforms other algorithms in terms of SINR-out in trials 1 to 4 and trials 7 to 9, and gives comparable performance in trial 5 and inferior performance in trial 6. It is noted that, in trials 4 to 6, there is no spatial interference considered (i.e., $s_i(t) = 0$). The performance for trial 5 indicates that the proposed NCFMV-EWPF is not as effective as the ZPF in suppressing the additive noise with higher SNR (SNR > 10 dB) when spatial interference is not present. Therefore, these experimental results demonstrate the superior capability of the proposed NCFMV-EWPF in suppressing the spatial and adverse additive interferences. For visualization purposes, the results in Table 1 have also been plotted in Figure 4, where the x-axis represents the SINR of the signal captured by the AVS (termed as SINR-input) and the y-axis represents the SINR of the enhanced speech (termed as SINR-out).

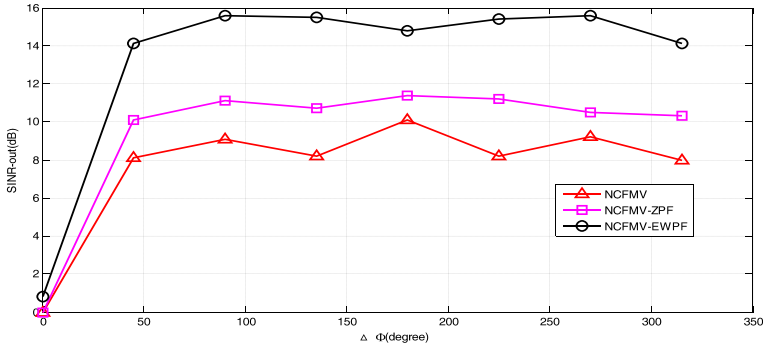

Figure 5 The SINR-out versus Δφ (NCFMV [5]).

Table 2 Performance comparison

	LSD	PESQ	SINR (dB)
INPUT	2.64	2.04	0.02
ZPF [4]	1.90	2.24	3.09
NCFMV [5]	2.55	2.29	8.12
NCFMV-ZPF [6]	*1.84*	*2.52*	11.94
NCFMV-EWPF (our proposed)	*1.84*	2.50	*13.85*

Results from the best performing methods are italicized.

5.1.2 Experiment 2: the impact of the angle between the target and interference speakers

This experiment evaluates the impact of the angle between the target and interference speakers ($\Delta\phi = \phi_s - \phi_i$) on the performance of the NCFMV-EWPF algorithm. The results of the SINR-out versus $\Delta\phi$ are shown in Figure 5, where the same experimental settings as those used for trial 7 in experiment 1 were adopted except the target speech location ϕ_s varied from (90°,0°) to (90°,360°) with 45° increments. From Figure 5, it is clear to see that when $\Delta\phi \to 0°$ (the target speaker moves closer to the interference speaker), for both algorithms, the SINR-out drops significantly and almost goes to 0. This means the speech enhancement is very much limited under this condition. However, when $\Delta\phi > 0°$, the SINR-out gradually increases. It is quite encouraging to see that the SINR-out of our proposed NCFMV-EWPF algorithm is superior to that of the NCFMV algorithm for all angles. Moreover, the SINR-out of our proposed NCFMV-EWPF algorithm maintains about 15 dB when $\Delta\phi \geq 45°$.

5.1.3 Experiment 3: SINR, LSD, and PESQ performance

In this experiment, we adopted three performance metrics (SINR, LSD, and PESQ) to evaluate the performance of the algorithms. The same experimental settings of those used in experiment 1 were employed, where the SINR-input is set as 0 dB, the target speaker is located at (90°, 45°), and the interference speaker is at (90°,0°) ($\Delta\phi = 45°$). The experimental results are given in Table 2. It can be seen that the overall performance of our proposed NCFMV-EWPF algorithm is superior to that of other comparison algorithms. The LSD and PESQ performance of the NCFMV-EWPF algorithm is comparable to that of the NCFMV-ZPF [6] algorithm. It is encouraging to see that the proposed NCFMV-EWPF algorithm is able to effectively suppress the interference and additive noise while maintaining good speech quality and less distortion.

5.2 Experiments on recorded data in an anechoic chamber

5.2.1 Experiment 4: the SINR-out performance with different speakers

In this experiment, we conducted the speech enhancement by using the recorded data from Ritz's lab [19]. The experimental setup is shown in Figure 6. The AVS has been built by two Knowles NR-3158 pressure gradient sensors (*u*-sensor and *v*-sensor) and one Knowles EK-3132 sensor (*o*-sensor) (Knowles Electronics Inc., Itasca, IL, USA). Recordings were made of 10 different speech sentences from the IEEE speech corpus [20] in an anechoic chamber and background noise only from computer servers and air conditioning. The anechoic chamber is similar to the noise field: $\boldsymbol{n}_{\text{avs}}(t) = 0$ and $s_i(t) \neq 0$. The sampling rate was 48 kHz and then down-sampled to 16 kHz for speech enhancement. The speakers were placed in front of the AVS at a distance of 1 m. Target speech was located at a fixed position (90°, 45°), while interference speech was located at (90°, 90°). Ten trials were carried out using the 10 different target speeches.

The experimental results are shown in Figure 7. The *x*-axis represents the number of trials, and the *y*-axis

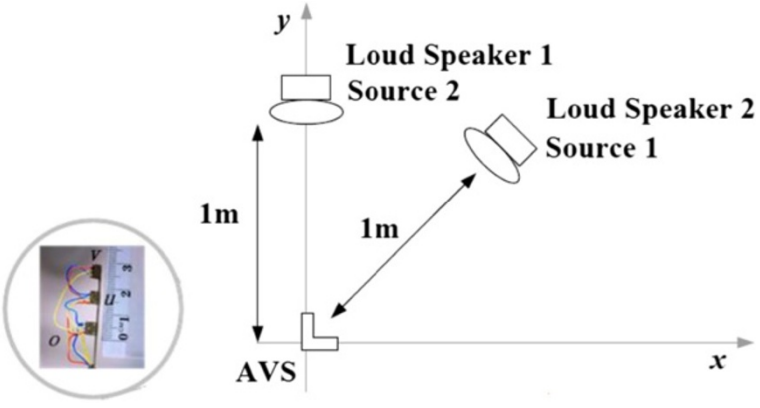

Figure 6 The illustration of the experimental setup in an anechoic chamber in Ritz's lab [19]. The leftmost image is the AVS used, $s(t)$ is the target speech (source 1) at 45°, and $s_i(t)$ is the interference speech (source 2) at 90°.

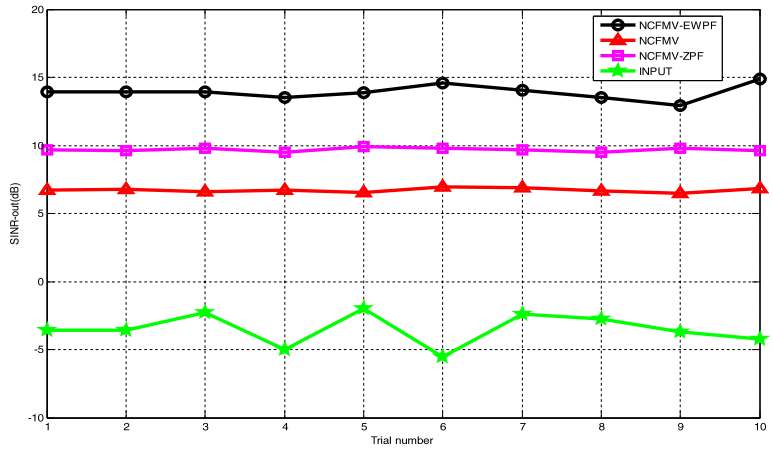

Figure 7 The SINR-out performance versus trial number (NCFMV [5]).

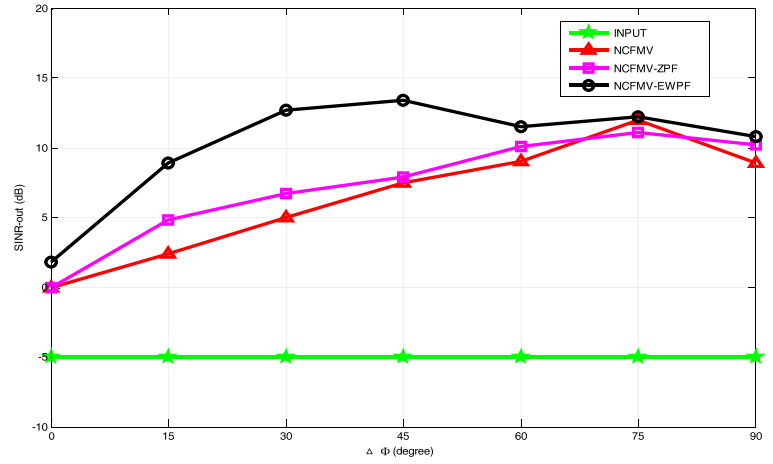

Figure 8 The SINR-out versus Δφ (NCFMV [5]).

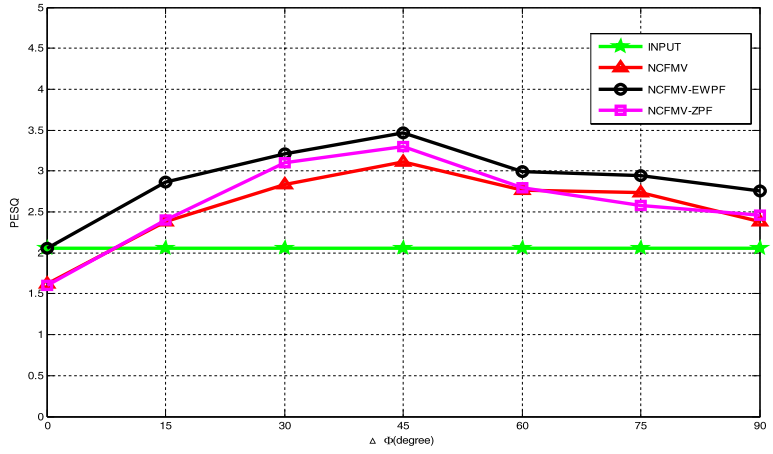

Figure 9 The PESQ performance versus Δφ (NCFMV [5]).

represents the SINR of the enhanced speech (in dB). It is clear to see that the proposed NCFMV-EWPF algorithm provides superior SINR-out performance for all trails when the SINR-input of the recorded data is at about −5 dB. The experimental results with the real recorded data further validate the effectiveness of the proposed NCFMV-EWPF in suppressing the strong competing speech.

5.2.2 Experiment 5: the impact of the angle between the target and interference speakers

Similar to experiment 2, this experiment evaluates the impact of the angle between the target and interference speakers ($\Delta\phi = |\phi_s - \phi_i|$) on the performance of the NCFMV-EWPF algorithm. The results of the SINR-out versus $\Delta\phi$ are shown in Figure 8, where the experimental setup is the same as that of experiment 4 except that the angle of the target speaker (ϕ_s) varies from (90°,90°) to (90°,0°) with 15° decrement.

From Figure 8, it is clear to see that the performance of the proposed NCFMV-EWPF algorithm is superior to that of the NCFMV algorithm for all $\Delta\phi$ values. Compared to the results shown in Figure 5 using the simulated data, similar conclusions can be drawn for the proposed NCFMV-EWPF algorithm. More specifically, with the recorded data, when $\Delta\phi > 15°$, the proposed NCFMV-EWPF algorithm can effectively enhance the target speech.

5.2.3 Experiment 6: PESQ performance versus $\Delta\varphi$

In this experiment, we only adopted one performance metrics (PESQ) to evaluate the performance of the algorithms. The same experimental settings as those used in experiment 5 were employed, where the angle of the interference speaker (ϕ_i) was fixed at (90°,90°) and the angle of the target speaker (ϕ_s) varied from (90°,90°) to (90°,0°) with 15° decrement. The experimental results are given in Figure 9. It can be seen that the overall performance of PESQ for our proposed NCFMV-EWPF algorithm is superior to that of the comparison algorithm for all angle differences. This experiment also demonstrates the ability of the proposed NCFMV-EWPF algorithm in effectively suppressing the interference and additive noise while maintaining good speech quality and less distortion when $\Delta\phi > 15°$.

6 Conclusions

In this paper, a novel speech enhancement algorithm named as NCFMV-EWPF has been derived with a single AVS unit by an efficient closed-form estimation of the power spectral densities of signals. The results of computer simulation show that the proposed NCFMV-EWPF algorithm outperforms the existing ZPF, NCFMV, and NCFMV-ZPF algorithms, in terms of suppressing the competing speaker and noise field. The results of real experiments show that compared with the NCFMV algorithms, the proposed NCFMV-EWPF algorithm can effectively suppress the competing speech and additive noise while maintaining good speech quality and less distortion. In addition, it is noted that the NCFMV-EWPF algorithm does not require the VAD technique, which not only reduces the computational complexity but also provides more robust performance in a noisy environment, such as the higher output SINR, less speech distortion, and better speech intelligibility. It is expected that this novel approach developed in this paper is a suitable solution for implementation within hands-free speech recording systems.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This work is partially supported by the National Natural Science Foundation of China (No. 61271309) and the Shenzhen Science & Technology Fundamental Research Program (No. JCY201110006). It was also partially supported by the Australian Research Council Grant DP1094053.

Author details

[1]ADSPLAB/ELIP, School of Electronic Computer Engineering, Peking University, Shenzhen 518055, China. [2]School of Electrical, Computer, and Telecommunications Engineering, University of Wollongong, Wollongong 2522, Australia.

Received: 21 January 2014 Accepted: 3 April 2014
Published: 27 April 2014

References

1. S Boll, Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process **27**(2), 113–120 (1979)
2. LJ Griffiths, CW Jim, An alternative approach to linearly constrained adaptive beamforming. IEEE Trans. Antennas Propag. **30**(1), 27–34 (1982)
3. YX Zou, SC Chan, B Wan, J Zhao, *Recursive robust variable loading MVDR beamforming in impulsive noise environment*, vol. 1–4 (Paper presented at the IEEE ASIA Pacific conference on circuits and system, Macao, 2008), pp. 988–991
4. R Zelinski, *A microphone array with adaptive post-filtering for noise reduction in reverberant rooms* (Paper presented at the IEEE international conference on acoustics, speech, and signal processing (ICASSP), New York, 1988)
5. ME Lockwood, DL Jones, Beamformer performance with acoustic vector sensors in air. J. Acoust. Soc. Am. **119**, 608–619 (2006)
6. IA McCowan, H Bourlard, Microphone array post-filter based on noise field coherence. IEEE Trans. Speech Audio Process **11**(6), 709–716 (2003)
7. J Benesty, MM Sondhi, Y Huang, *Springer Handbook of Speech Processing* (Springer, Berlin-Heidelberg, 2008)
8. SV Vaseghi, *Advanced Digital Signal Processing and Noise Reduction*, 2nd edn. (John Wiley & Sons ltd, Chichester, 2000)
9. J Bitzer, KU Simmer, KD Kammeyer, *Multichannel noise reduction algorithms and theoretical limits* (Paper presented at EURASIP European signal processing conference (EUSIPCO), Rhodes, 1998)
10. ME Lockwood, DL Jones, RC Bilger, CR Lansing, WDO Brien, BC Wheeler, AS Feng, Performance of time-and frequency-domain binaural beamformers based on recorded signals from real rooms. J. Acoust. Soc. Am. **115**, 379 (2004)
11. M Shujau, CH Ritz, IS Burnett, *Speech enhancement via separation of sources from co-located microphone recordings* (Paper presented at IEEE international conference on acoustics, speech and signal processing (ICASSAP), Dallas, 2010)
12. PKT Wu, C Jin, A Kan, *A multi-microphone speech enhancement algorithm tested using acoustic vector sensor* (Paper presented at the 12th international workshop on acoustic echo and noise control, Tel-Aviv-Jaffa, 2010)
13. B Li, YX Zou, *Improved DOA estimation with acoustic vector sensor arrays using spatial sparsity and subarray manifold* (Paper presented at IEEE

international conference on acoustics, speech and signal processing (ICASSP), Kyoto, 2012)
14. W Shi, YX Zou, B Li, CH Ritz, M Shujau, J Xi, *Multisource DOA estimation based on time-frequency sparsity and joint inter-sensor data ratio with single acoustic vector sensor* (Paper presented at IEEE international conference on acoustics, speech and signal processing (ICASSP), Vancouver, 2013)
15. M Shujau, *In air acoustic vector sensors for capturing and processing of speech signals, Dissertation* (University of Wollongong, , 2011)
16. R Gray, A Buzo, JA Gray, Y Matsuyama, Distortion measures for speech processing. IEEE Trans. Acoust. Speech Signal Process **28**(4), 367–376 (1980)
17. ITU-T, *Recommendation P.862 - Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs* (International Telecommunication Union - Telecommunication Standardization Sector, Geneva, 2001)
18. NOISEX-92: http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html
19. CH Ritz, IS Burnett, *Separation of speech sources using an acoustic vector sensor* (Paper presented at IEEE international workshop on multimedia signal processing, Hanzhou, 2011)
20. IEEE Subcommittee, IEEE recommended practice for speech quality measurements. IEEE Trans. Audio Electro-acoustics **AU-17**(3), 225–246 (1969)

doi:10.1186/1687-4722-2014-17
Cite this article as: Zou *et al.*: Speech enhancement with an acoustic vector sensor: an effective adaptive beamforming and post-filtering approach. *EURASIP Journal on Audio, Speech, and Music Processing* 2014 **2014**:17.