Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2008, Article ID 860360, 10 pages
doi:10.1155/2008/860360

Research Article

Postfiltering Using Multichannel Spectral Estimation in Multispeaker Environments

Hai Quang Dam, Sven Nordholm, Hai Huyen Dam, and Siow Yong Low

Western Australian Telecommunications Research Institute (WATRI), Crawley, WA 6009, Australia

Correspondence should be addressed to Hai Quang Dam, amhai@watri.org.au

Received 14 September 2006; Accepted 5 July 2007

Recommended by Douglas O'Shaughnessy

This paper investigates the problem of enhancing a single desired speech source from a mixture of signals in multispeaker environments. A beamformer structure is proposed which combines a fixed beamformer with postfiltering. In the first stage, the fixed multiobjective optimal beamformer is designed to spatially extract the desired source by suppressing all other undesired sources. In the second stage, a multichannel power spectral estimator is proposed and incorporated in the postfilter, thus enabling further suppression capability. The combined scheme exploits both spatial and spectral characteristics of the signals. Two new multichannel spectral estimation methods are proposed for the postfiltering using, respectively, inner product and joint diagonalization. Evaluations using recordings from a real-room environment show that the proposed beamformer offers a good interference suppression level whilst maintaining a low-distortion level of the desired source.

Copyright © 2008 Hai Quang Dam et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. INTRODUCTION

Multichannel beamforming techniques can be largely divided into three types, namely, fixed, optimum, and adaptive beamforming [1, 2]. For a fixed beamformer, the beamformer weights, which usually consist of FIR-filter weights, are designed to focus into a main source direction while suppressing signals from other undesired directions. This problem can be viewed as a multidimensional filter design problem [2]. As such, the weights are calculated based on information about the array geometry and the source localization with no statistical information about the signal's environment or the required signals.

Multichannel optimum filtering, on the other hand, requires statistical knowledge about the noise statistics, the environment, and the source statistics. The beamformer coefficients are optimized in such a manner that a focussed beam is steered to a desired source direction, whilst suppressing the contributions coming from other directions [2, 3]. Similar to the fixed beamformer case, the design also requires information about the location of the target signal and the array geometry. From those parameters, a spatial, spectral, and temporal filter is formed to match the beamforming requirement [4, 5].

Adaptive beamforming techniques are developed to track time-varying signal situations [6, 7]. A well-known technique is to combine the beamformer with an adaptive postfiltering technique. The adaptive postfiltering uses the estimation of spectral densities of the desired and undesired signals in the filter output to further suppress the noise. One common method to perform postfiltering is spectral subtraction. This method exploits spectral information of the noise and the speech sources to form a gain function to suppress the noise [8, 9]. A critical part for spectral subtraction is the detection of speech active and inactive periods [10]. The speech inactive periods are used to update the noise statistics. During these periods, the noise information is updated in the gain function. Naturally, any misdetection will lead to erroneous update of the noise and result in distortion. Also, spectral subtraction succumbs to nonstationary noise as it relies heavily on speech pauses to update the noise statistics. More explicitly, the noise is estimated during speech pauses and is used to form the gain function during speech periods. As a consequence, spectral subtraction cannot deal well with situations where the interference is another speech source or the noise is nonstationary.

To resolve the nonstationary problem, Zelinski introduced the multichannel postfiltering technique [11]. The

postfilter uses the auto- and cross-spectral densities of the array inputs to estimate the signal and noise spectral densities. By doing so, the postfilter is capable of performing in nonstationary noise. However, one of the main assumptions in [11] is that the noise in different channels are uncorrelated corresponding to an incoherent noise field. In practice, the correlation of the noise signals between channels may be significant. This is especially the case for closely spaced sensors, for example, typically in speech enhancement applications. To cope with that, a number of techniques have been proposed during the past few years [12, 13]. A postfiltering technique based on the complex coherence function for a specific coherence noise field such as spherically isotropic (diffuse) or cylindrically isotropic noise fields is proposed in [12]. In [13], a multichannel postfiltering is developed to minimize the log-spectral amplitude distortion in nonstationary noise environments. A main assumption is made that a desired source component is stronger at the beamformer output than at any reference noisy signal, and the interference component is the strongest at one of the reference signals. However, this assumption might not be satisfied if the desired source and other undesired interferences are located close to the array and have fast time-varying characteristics such as speech signals.

This paper aims to recover a particular speech source while rejecting other speech sources in multispeaker environments. This has been referred to as a cocktail party effect or an "attentional selectivity" [14, 15]. As an example, consider a situation with many speakers in a "meeting" room. The observed signals contain the speech signals from many speakers with the possibility of overlapping one another. The objective is to extract a single desired signal from the mixtures.

A new beamformer structure is proposed which employs a multichannel power spectral estimator of the desired speech source. This structure includes a multiobjective optimal beamformer followed by a postfilter. The multiobjective optimal beamformer is designed to spatially extract a desired source while suppressing all other undesired source(s). More specifically, if there are three or more speech sources, the multiobjective optimal beamformer is designed to eliminate at least two undesired sources. As such, it may not be able to suppress all the undesired sources. To suppress further the undesired sources from the beamformer output, an adaptive postfilter is proposed which includes a multichannel spectral estimation of the desired signal. Two multichannel spectral estimation methods are developed for the postfiltering using, respectively, inner product and joint diagonalization to estimate the desired source power spectral density (PSD). Evaluations using recordings from a real room environment show that the proposed beamformers offer good interference suppression levels whilst maintaining low distortion levels of the desired source.

The organization of the paper is given as follows. The problem formulation is outlined in Section 2. The spatial correlation matrix estimation using calibration signals is developed in Section 3. A fixed multiobjective optimal beamformer is proposed in Section 4. Two multichannel spectral estimation methods using, respectively, inner product and joint diagonalization are developed in Section 5. Finally, evaluations of the proposed beamformer using real data are presented in Section 6, and conclusions are given in Section 7.

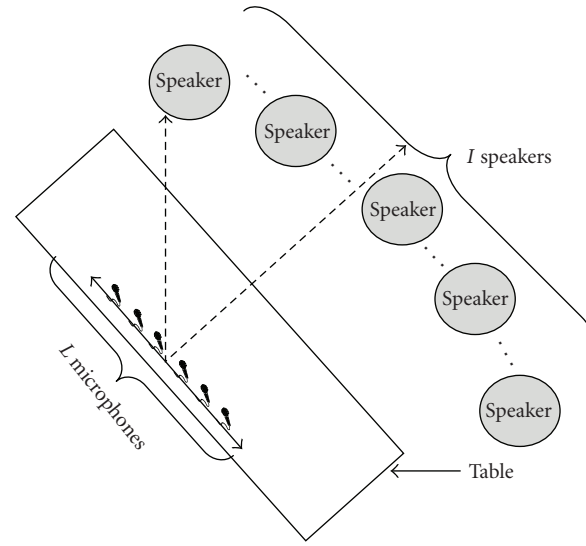

Figure 1: Position of sources and the microphone array in multispeaker environment.

2. PROBLEM FORMULATION

Consider a multispeaker situation with I speakers located in the near field of an L-element microphone array as depicted in Figure 1. The speakers can be active in a random manner and their speech signals may overlap in time. Denote by $\mathbf{s}_i(n)$, $1 \le i \le I$, an $L \times 1$ vector of the discrete-time observed signal from the ith source at the microphones where n denotes the time index. The received signal $\mathbf{x}(n)$ at the microphones can be written as

$$\mathbf{x}(n) = \sum_{i=1}^{I} \mathbf{s}_i(n) + \mathbf{v}(n), \tag{1}$$

where $\mathbf{v}(n)$ is the background noise. Here, we concentrate mainly on the case with speech mixtures. Thus, the term $\mathbf{v}(n)$ is being omitted. The task at hand is to extract the desired source(s) from a mixture of I sources.

The proposed beamformer is performed in the frequency domain. Thus, the received signal is decomposed into M subbands in the frequency domain by using an analysis filter bank [16]. The filtering and processing are then performed for each frequency bin. The observed signal $\mathbf{x}(\omega, k)$ for each frequency bin ω and time index k can be given as

$$\mathbf{x}(\omega, k) = \sum_{i=1}^{I} \mathbf{s}_i(\omega, k), \tag{2}$$

where $\mathbf{s}_i(\omega, k)$ is the contribution from the ith source. Denote by $\overline{\mathbf{R}}_i(\omega)$ and $p_i(\omega, k)$ the spatial correlation matrix and the PSD at time instant k, respectively, of the ith source [17, 18]. By assuming that all the sources are spatially invariant and

statistically independent, the correlation matrix of the received signal $\mathbf{R}_x(\omega, k)$ at instant k can be expressed as

$$\mathbf{R}_x(\omega, k) = \sum_{i=1}^{I} \overline{\mathbf{R}}_i(\omega) p_i(\omega, k). \tag{3}$$

In the following section, a calibration method will be presented to calculate the source spatial correlation matrices before the beamforming process.

3. SPATIAL CORRELATION MATRIX ESTIMATION USING CALIBRATION SIGNALS

In [19, 20], a calibration method is outlined where the training samples of the sources are recorded prior to the beamforming process. This method is developed to estimate the statistical information of the sources which includes unknown signal path information. By doing so, all the information on the array geometry and source localization will be reflected in the solution [21].

During the calibration period, each speaker is active for a short period of time while other speakers are silent. Denote by $[K_{1,i}, K_{2,i}]$ the active time of the ith source and $\hat{\mathbf{R}}_{i,\text{cal}}(\omega)$ the correlation matrix for ith source estimated during the calibration period. This matrix can be obtained as

$$\hat{\mathbf{R}}_{i,\text{cal}}(\omega) = \frac{1}{K_{2,i} - K_{1,i} + 1} \sum_{k=K_{1,i}}^{K_{2,i}} \mathbf{x}(\omega, k)\mathbf{x}^H(\omega, k). \tag{4}$$

Moreover, denote by $\hat{\mathbf{d}}_{i,\text{cal}}(\omega)$ the spatial cross correlation vector with respect to the ℓth prechosen reference microphone, $1 \le \ell \le L$. The vector $\hat{\mathbf{d}}_{i,\text{cal}}(\omega)$ is estimated as

$$\hat{\mathbf{d}}_{i,\text{cal}}(\omega) = \frac{1}{K_{2,i} - K_{1,i} + 1} \sum_{k=K_{1,i}}^{K_{2,i}} \mathbf{x}(\omega, k)x^*(\omega, k, \ell), \tag{5}$$

where $x(\omega, k, \ell)$ is the received signal at the ℓth microphone. The spatial correlation matrix $\overline{\mathbf{R}}_i(\omega)$ and the spatial cross correlation vector $\overline{\mathbf{d}}_i(\omega)$ can be estimated as

$$\overline{\mathbf{R}}_i(\omega) = \frac{\hat{\mathbf{R}}_{i,\text{cal}}(\omega)}{\hat{R}_{i,\text{cal}}(\omega, \ell, \ell)}, \tag{6}$$

$$\overline{\mathbf{d}}_i(\omega) = \frac{\hat{\mathbf{d}}_{i,\text{cal}}(\omega)}{\hat{d}_{i,\text{cal}}(\omega, \ell)}, \tag{7}$$

where $\hat{\mathbf{R}}_{i,\text{cal}}(\omega, \ell, \ell)$ is the (ℓ, ℓ) element of the matrix $\hat{\mathbf{R}}_{i,\text{cal}}(\omega)$ and $\hat{d}_{i,\text{cal}}(\omega, \ell)$ is the ℓth element of the vector $\hat{\mathbf{d}}_{i,\text{cal}}(\omega)$. Next, a fixed multiobjective optimal beamformer is developed utilizing the spatial correlation matrices.

4. FIXED MULTIOBJECTIVE OPTIMAL BEAMFORMER

In this section, a fixed multiobjective optimal beamformer incorporating the spatial correlation matrices is proposed to suppress the interference signals whilst preserving the desired speech. For simplicity, the first source $\mathbf{s}_1(\omega, k)$ is assumed to be the desired source while other $I - 1$ sources, $\mathbf{s}_i(\omega, k)$, $2 \le i \le I$, are undesired. The fixed multiobjective optimal filter weight $\mathbf{w}_f(\omega)$ for the frequency ω is designed to minimize

$$\mathbf{w}_f^H(\omega)\overline{\mathbf{R}}_i(\omega)\mathbf{w}_f(\omega) \quad \forall 2 \le i \le I, \tag{8}$$

while maintaining the desired source direction, for example, the first source direction

$$\mathbf{w}_f^H(\omega)\overline{\mathbf{d}}_1(\omega) = 1. \tag{9}$$

Thus, we propose to minimize the following weighted cost function:

$$J = \mathbf{w}_f^H(\omega)\left[\sum_{i=2}^{I} \overline{\mathbf{R}}_i(\omega)\gamma_i(\omega)\right]\mathbf{w}_f(\omega), \tag{10}$$

where $\gamma_i(\omega)$, $2 \le i \le I$, are the weighting parameters for the sources. One possibility is to choose $\gamma_i(\omega)$ as the calibration values $\hat{\mathbf{R}}_{i,\text{cal}}(\omega, \ell, \ell)$ in (6) to match the spectral proportion among the sources in the calibration time. Another possibility is to choose $\gamma_i(\omega)$ as one to give equal weighting for all interference sources. In general, $\gamma_i(\omega)$ can be chosen differently to allow different suppression levels for the interference depending on the requirements. Consequently, the fixed multiobjective optimal beamformer weight can be obtained by solving the following optimization problem:

$$\begin{aligned} &\min_{\mathbf{w}(\omega)} \mathbf{w}^H(\omega)\left[\sum_{i=2}^{I} \overline{\mathbf{R}}_i(\omega)\gamma_i(\omega)\right]\mathbf{w}(\omega) \\ &\text{subject to } \mathbf{w}^H(\omega)\overline{\mathbf{d}}_1(\omega) = 1. \end{aligned} \tag{11}$$

The solution of this optimization problem can be expressed as

$$\mathbf{w}_f(\omega) = \frac{\left[\sum_{i=2}^{I} \overline{\mathbf{R}}_i(\omega)\gamma_i(\omega)\right]^{-1}\overline{\mathbf{d}}_1(\omega)}{\overline{\mathbf{d}}_1^H(\omega)\left[\sum_{i=2}^{I} \overline{\mathbf{R}}_i(\omega)\gamma_i(\omega)\right]^{-1}\overline{\mathbf{d}}_1(\omega)}. \tag{12}$$

The output of the fixed beamformer is calculated as

$$u(\omega, k) = \mathbf{w}_f^H(\omega)\mathbf{x}(\omega, k). \tag{13}$$

The beamformer output is then passed through a postfilter to further suppress the undesired signals.

5. POSTFILTERING USING MULTICHANNEL SPECTRAL ESTIMATION

In this section, a postfiltering method employing two new multichannel spectral estimators is proposed to suppress further the undesired sources in the fixed multiobjective optimal beamformer output while maintaining the desired source component. More specifically, the spatial difference between the desired and the undesired sources is used for the PSD estimation of the desired source.

To track the spectral changes of the desired speech source, the multichannel spectral estimator is performed in the periods where the speech sources are quasistationary. As such, at a time instant k, the instantaneous PSD of the desired source

is estimated based on K samples before this instant. The estimated correlation matrix $\hat{\mathbf{R}}_x(\omega,k)$ of the observed signals for these K samples is calculated as

$$\hat{\mathbf{R}}_x(\omega,k)=\frac{1}{K+1}\sum_{n=k-K}^{k}\mathbf{x}(\omega,n)\mathbf{x}^H(\omega,n). \tag{14}$$

Since speech sources can assume to be spatially invariant during the period of K consecutive samples, the model in (3) is employed. Based on (3) and (14), we propose two different multichannel spectral estimators to efficiently estimate the desired source PSD, $p_1(\omega,k)$, from a mixture of signals in multispeaker environments.

5.1. Spectral estimation using an inner product and determinant

A PSD estimation method of the desired source, $p_1(\omega,k)$, is proposed based on the estimated instantaneous correlation matrix $\hat{\mathbf{R}}_x(\omega,k)$ and the model of the instantaneous correlation matrix given in (3). Since the spatial correlation matrices $\overline{\mathbf{R}}_i(\omega)$, $1\le i\le I$, are known from calibration and $\hat{\mathbf{R}}_x(\omega,k)$ has been estimated, the task is to find $p_1(\omega,k)$ or in a more general case $p_i(\omega,k)$. This method relies on properties of determinants and full rank matrices.

For every calibration matrix $\overline{\mathbf{R}}_i(\omega)$ of size $L\times L$, define an $2L^2\times 1$ real vector $V\{\overline{\mathbf{R}}_i(\omega)\}$ containing all the elements of $\overline{\mathbf{R}}_i(\omega)$ as

$$V\{\overline{\mathbf{R}}_i(\omega)\}=\left[(\mathbf{r}_1^{\Re})^T,(\mathbf{r}_1^{\Im})^T,(\mathbf{r}_2^{\Re})^T,(\mathbf{r}_2^{\Im})^T,\ldots,(\mathbf{r}_L^{\Re})^T,(\mathbf{r}_L^{\Im})^T\right]^T, \tag{15}$$

where $\mathbf{r}_l^{\Re}$ and $\mathbf{r}_l^{\Im}$ are, respectively, the real and imaginary parts of the lth column of $\overline{\mathbf{R}}_i(\omega)$ for all $1\le l\le L$. Using the vectors $V\{\overline{\mathbf{R}}_i(\omega)\}$, we form a matrix $\Gamma(\omega)$ as

$$\Gamma(\omega)=\begin{pmatrix}\zeta(1,1) & \zeta(1,2) & \cdots & \zeta(1,I)\\ \zeta(2,1) & \zeta(2,2) & \cdots & \zeta(2,I)\\ \vdots & \vdots & \ddots & \vdots\\ \zeta(I,1) & \zeta(I,2) & \cdots & \zeta(I,I)\end{pmatrix}, \tag{16}$$

where $\zeta(i,j)$, $1\le i,j\le I$, is the inner product between $V\{\overline{\mathbf{R}}_i(\omega)\}$ and $V\{\overline{\mathbf{R}}_j(\omega)\}$:

$$\zeta(i,j)=\frac{1}{2L^2}V^T\{\overline{\mathbf{R}}_i(\omega)\}V\{\overline{\mathbf{R}}_j(\omega)\}. \tag{17}$$

Since $\overline{\mathbf{R}}_i(\omega)$, $1\le i\le I$, are spatial correlation matrices of the speech sources with strictly different locations, their corresponding vectors can assume to be linearly independent. From this, it follows that the determinant of the matrix $\Gamma(\omega)$, denoted by $\det\{\Gamma(\omega)\}$, is nonzero [22].

In the same way as in (15), a vector $V\{\mathbf{R}_x(\omega,k)\}$ can be formed from $\mathbf{R}_x(\omega,k)$. Since the operation from $\mathbf{R}_x(\omega,k)$ to $V\{\mathbf{R}_x(\omega,k)\}$ is linear, by using (21) the following expression is obtained:

$$V\{\mathbf{R}_x(\omega,k)\}=\sum_{i=1}^{I}p_i(\omega,k)V\{\overline{\mathbf{R}}_i(\omega)\}. \tag{18}$$

Inserting this expression in (17) yields

$$\begin{aligned}\zeta_x(i)&=\sum_{j=1}^{I}\zeta(j,i)p_j(\omega,k)\\&=\zeta(1,i)p_1(\omega,k)+\sum_{j=2}^{I}\zeta(j,i)p_j(\omega,k),\end{aligned} \tag{19}$$

where $\zeta_x(i)$, $1\le i\le I$, is the inner product between the instantaneous correlation matrix $\mathbf{R}_x(\omega,k)$ and the spatial correlation matrices $\overline{\mathbf{R}}_i(\omega)$. Inserting $\zeta_x(i)$, $1\le i\le I$, in the first row of the matrix $\Gamma(\omega,k)$ in (16), we have

$$\Gamma_x(\omega,k)=\begin{pmatrix}\zeta_x(1) & \zeta_x(2) & \cdots & \zeta_x(I)\\ \zeta(2,1) & \zeta(2,2) & \cdots & \zeta(2,I)\\ \vdots & \vdots & \ddots & \vdots\\ \zeta(I,1) & \zeta(I,2) & \cdots & \zeta(I,I)\end{pmatrix}. \tag{20}$$

By combining (19) and (20), we have (21).

$$\begin{aligned}&\Gamma_x(\omega,k)\\&=\begin{pmatrix}p_1(\omega,k)\zeta(1,1) & p_1(\omega,k)\zeta(1,2) & \cdots & p_1(\omega,k)\zeta(1,I)\\ \zeta(2,1) & \zeta(2,2) & \cdots & \zeta(2,I)\\ \vdots & \vdots & \ddots & \vdots\\ \zeta(I,1) & \zeta(I,2) & \cdots & \zeta(I,I)\end{pmatrix}\\&+\begin{pmatrix}\sum_{j=2}^{I}\zeta(j,1)p_j(\omega,k) & \sum_{j=2}^{I}\zeta(j,2)p_j(\omega,k) & \cdots & \sum_{j=2}^{I}\zeta(j,I)p_j(\omega,k)\\ 0 & 0 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 0\end{pmatrix}.\end{aligned} \tag{21}$$

By taking the determinant of (21), we have

$$\det\{\Gamma_x(\omega,k)\}=p_1(\omega,k)\det\{\Gamma(\omega)\}. \tag{22}$$

Thus, we propose an estimation method for $p_1(\omega,k)$ based on $\det\{\Gamma(\omega)\}$ and $\det\{\Gamma_{\hat{x}}(\omega,k)\}$ as

$$p_1(\omega,k)=\max\left\{0,\frac{\det\{\Gamma_{\hat{x}}(\omega,k)\}}{\det\{\Gamma(\omega)\}}\right\}, \tag{23}$$

where $\Gamma_{\hat{x}}(\omega,k)$ is the same as $\Gamma_x(\omega,k)$ but with $\mathbf{R}_x(\omega,k)$ replaced by the estimate of the correlation matrix $\hat{\mathbf{R}}_x(\omega,k)$.

It can be noted from (20) that for each time instant k, we only need to estimate the first row of the matrix $\Gamma_x(\omega,k)$. This is done by taking the inner product between $V\{\hat{\mathbf{R}}_x(\omega,k)\}$ and $V\{\overline{\mathbf{R}}_i(\omega)\}$ for all i. As the matrices $V\{\overline{\mathbf{R}}_i(\omega)\}$ are all known, this results in $2IL^2$ real multiplications. In addition, the determinant $\det\{\Gamma_x(\omega,k)\}$ in (23) requires I real multiplications where the determinant is taken along the first row with all the cofactors precalculated. Therefore, the number of real multiplications required is approximately $I(2L^2+1)$ for each frequency bin.

In the following section, we present another method for estimating the desired source PSD by using a joint diagonalization technique.

5.2. Spectral estimation using joint diagonalization

Since the spatial correlation matrices of all the undesired sources are known, joint diagonalization is proposed to be performed prior to the beamforming period to extract information of the undesired signals. As such, for each frequency bin ω, the problem becomes to estimate the matrix $\mathbf{H}(\omega)$ which jointly minimizes the off-diagonal elements of the following matrices:

$$\mathbf{H}(\omega)\overline{\mathbf{R}}_2(\omega)\mathbf{H}^H(\omega),\ldots,\mathbf{H}(\omega)\overline{\mathbf{R}}_i(\omega)\mathbf{H}^H(\omega). \tag{24}$$

To avoid trivial solutions, the following constraint is included:

$$\left|\left|\mathbf{h}_i(\omega)\right|\right|_{\mathcal{F}} = 1, \quad 1 \le i \le L, \tag{25}$$

where $\mathbf{h}_i(\omega)$ is the ith column of the matrix $\mathbf{H}(\omega)$ and $\|\cdot\|_{\mathcal{F}}$ is the Frobenius norm operator. This problem can be formulated as minimizing the following cost function:

$$C(\omega) = \sum_{i=2}^{I} \left|\left| \text{offdiag}\left\{\mathbf{H}(\omega)\overline{\mathbf{R}}_i(\omega)\mathbf{H}^H(\omega)\right\}\right|\right|_{\mathcal{F}}^2, \tag{26}$$

with the constraints in (25), where offdiag$\{\cdot\}$ is an operator that sets all diagonal elements of $\{\cdot\}$ to zeros. Here, this optimization problem is solved by using the algorithm proposed in [23], where the simultaneous diagonalization algorithm is an extension of the Jacobi technique, that is, a joint diagonality criterion is iteratively optimized under plane rotations.

Denote by $\mathbf{H}(\omega)$ the optimum solution for the joint diagonalization problem. The desired source PSD, $p_1(\omega,k)$, is estimated from the correlation matrix $\hat{\mathbf{R}}_x(\omega,k)$ of the observed signal and the matrix $\mathbf{H}(\omega)$ according to

$$\begin{aligned} p_1(\omega,k) &= \arg\min_{p_1(\omega)\ge 0} \left|\left| \text{offdiag}\left\{\mathbf{H}(\omega)\hat{\mathbf{R}}_x(\omega,k)\mathbf{H}^H(\omega)\right\}\right.\right. \\ &\qquad \left.\left. - p_1(\omega)\,\text{offdiag}\left\{\mathbf{H}(\omega)\overline{\mathbf{R}}_1(\omega)\mathbf{H}^H(\omega)\right\}\right|\right|_{\mathcal{F}}^2. \end{aligned} \tag{27}$$

Denote by $r_{m,n}(\omega,k)$, $h_{m,n}(\omega)$, $a_{m,n}(\omega,k)$, and $b_{m,n}(\omega)$ the (m,n)th complex elements of the matrices $\hat{\mathbf{R}}_x(\omega,k)$, $\mathbf{H}(\omega)$, $\mathbf{H}(\omega)\hat{\mathbf{R}}_x(\omega,k)\mathbf{H}^H(\omega)$, and $\mathbf{H}(\omega)\overline{\mathbf{R}}_1(\omega)\mathbf{H}^H(\omega)$, respectively. The element $a_{m,n}(\omega,k)$ can be obtained as

$$a_{m,n}(\omega,k) = \sum_{i=1}^{L}\sum_{j=1}^{L} h_{m,i}(\omega)r_{i,j}(\omega,k)h_{n,j}^*(\omega). \tag{28}$$

Since, the right-hand side of (27) is an algebraic polynomial of degree 2 with an unknown parameter $p_1(\omega)$, the optimization solution with constraint $p_1(\omega) \ge 0$ can be written as

$$p_1(\omega,k) = \max\left\{0, \frac{\sum_{\substack{m=1\\m\ne n}}^{L}\sum_{n=1}^{L}\Re\left\{a_{mn}(\omega,k)b_{mn}^*(\omega)\right\}}{\sum_{\substack{m=1\\m\ne n}}^{L}\sum_{n=1}^{L}\left|b_{mn}(\omega)\right|^2}\right\}, \tag{29}$$

where $\Re\{\cdot\}$ denotes the real part of a complex variable. Using (28), the term in the right-hand side of (29) can be written as

$$\begin{aligned} &\frac{\sum_{\substack{m=1\\m\ne n}}^{L}\sum_{n=1}^{L}\Re\left\{a_{mn}(\omega,k)b_{mn}^*(\omega)\right\}}{\sum_{\substack{m=1\\m\ne n}}^{L}\sum_{n=1}^{L}\left|b_{mn}(\omega)\right|^2} \\ &= \sum_{i=1}^{L}\sum_{j=1}^{L}\Re\left\{r_{i,j}(\omega,k)\frac{\sum_{\substack{m=1\\m\ne n}}^{L}\sum_{n=1}^{L}h_{m,i}(\omega)h_{n,j}^*(\omega)b_{mn}^*(\omega)}{\sum_{\substack{m=1\\m\ne n}}^{L}\sum_{n=1}^{L}\left|b_{mn}(\omega)\right|^2}\right\}. \end{aligned} \tag{30}$$

As such, the solution (29) can be obtained by multiplying the variables $r_{i,j}(\omega,k)$ with the precalculated cofactors. So, the number of calculations required for each estimation step is approximately L^2 complex multiplications or $4L^2$ real multiplications for each frequency bin.

The desired source PSD is now used in the postfilter to improve the performance of the fixed multiobjective optimal beamformer.

5.3. Postfilter

Since the first signal is assumed to be the desired source, the power of the desired source in the output of the fixed multiobjective optimal beamformer at a time instant k, $P_d(\omega,k)$, can be estimated as

$$P_d(\omega,k) = p_1(\omega,k)\mathbf{w}_f^H(\omega)\overline{\mathbf{R}}_1(\omega)\mathbf{w}_f(\omega). \tag{31}$$

The total power of the output, $P(\omega,k)$, can be estimated based on $\hat{\mathbf{R}}_x(\omega,k)$ as

$$P(\omega,k) = \mathbf{w}_f^H(\omega)\hat{\mathbf{R}}_x(\omega,k)\mathbf{w}_f(\omega). \tag{32}$$

From (31) and (32), the source power gain in the postfilter output can be calculated as

$$G(\omega,k) = \min\left(1, \sqrt{\frac{P_d(\omega,k)}{P(\omega,k)}}\right), \quad P(\omega,k) > 0. \tag{33}$$

If $P(\omega,k)$ is zero, then $G(\omega,k)$ is set to one to avoid a numerical problem. The output of the postfilter can be obtained based on the beamformer output $u(\omega,k)$ in (13) and the gain $G(\omega,k)$ as

$$y(\omega,k) = G(\omega,k)u(\omega,k). \tag{34}$$

This output is then passed through a synthesis filter bank to obtain its fullband representation [16]. A general diagram of the proposed structure is shown in Figure 2.

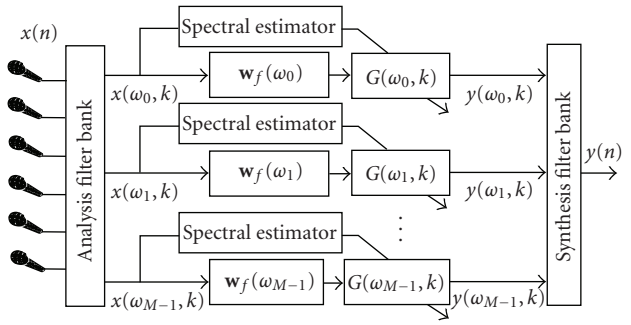

Figure 2: Multi-objective optimal beamforming with postfiltering using the analysis and synthesis filter banks.

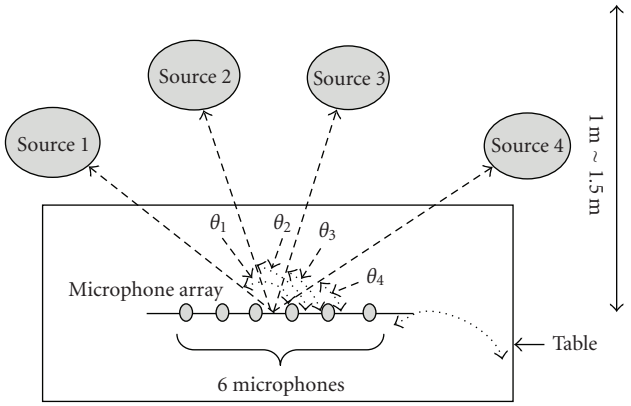

Figure 3: Position of original sources and the microphone array in the two-dimensional space.

6. EVALUATIONS

Measurements and evaluations have been performed in a real room environment using a linear microphone array consisting of 6 microphones with the distance of 6 cm between two adjacent microphones. There are 4 near-field speakers (2 men and 2 women). The distance between the speakers and the microphone array is approximately 1 m. The room size is $3.5 \times 3.1 \times 2.3\ \mathrm{m}^3$ with the reverberation time approximately 250 milliseconds. The speaker number is 1 to 4 from left to right.

The positions of the speakers are shown in Figure 3 with θ_1, θ_2, θ_3, and θ_4 being approximately 145°, 110°, 70°, and 35°, respectively.

The calibration time for each speaker is 10 seconds. This calibration time can be chosen arbitrarily. However, it is recommended that the calibration time is chosen more than 3 seconds to capture the spatial information of the speakers. The weighting parameters $\gamma_i(\omega)$ in (10) are chosen as $p_{i,\mathrm{cal}}(\omega)$.

Figure 4 shows the time domain plots of the speech signals and the observed signal at the 4th microphone. The length of the speaker speech signals is 35 seconds and the speech signals were recorded separately for the evaluations. Note that the recording was made from the actual human speakers and the speech signals occurred at different times

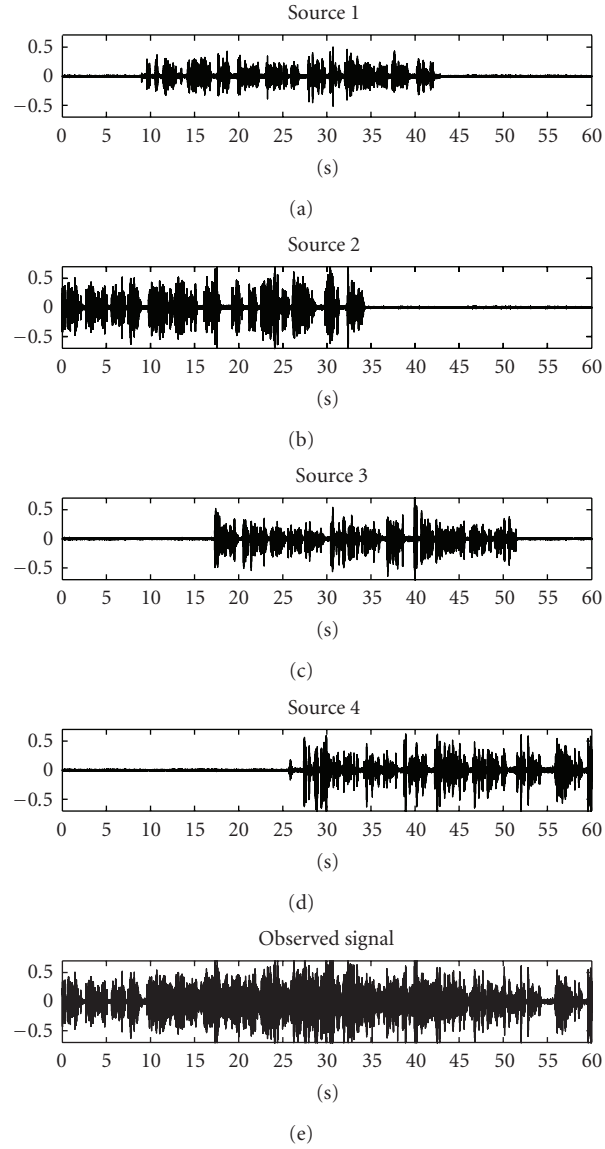

Figure 4: Time domain plots of the original sources and the observed signal at the 4th microphone.

and overlaped each other. The overlapping is used to simulate simultaneous conversation between the speakers. The corresponding spectrogram plots of the speech signals and the observed signal at the 4th microphone are depicted in Figure 5.

The observed signals are decomposed into $M = 64$ subbands by using a uniform oversampled analysis filterbank. In this case, a oversampling factor of two is chosen to reduce the aliasing effects between adjacent subbands [16]. The performance of the proposed beamformer is measured in terms of the interference suppression (IS) level, defined as

$$\mathrm{IS} = 10\log_{10}\left(\frac{\int_{-\pi}^{\pi}\widehat{P}_{\mathrm{in},n}(\omega)d\omega}{\int_{-\pi}^{\pi}\widehat{P}_{\mathrm{out},n}(\omega)d\omega}\right) - 10\log_{10}(C_d), \qquad (35)$$

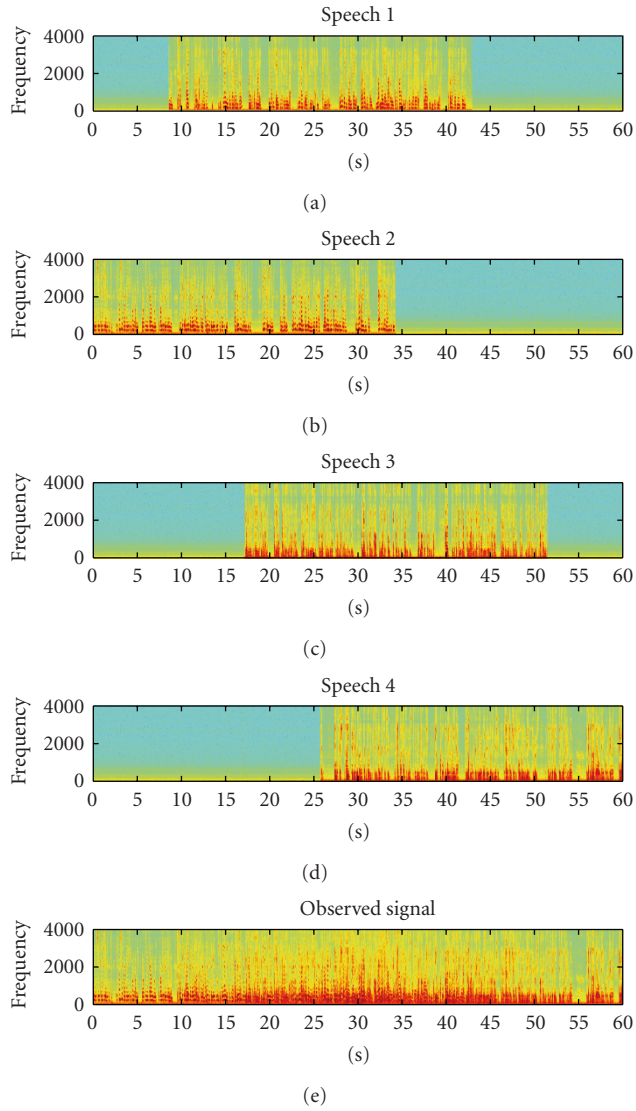

Figure 5: Spectrograms of the original sources and the observed signal at the 4th microphone.

where $\hat{P}_{\text{in},n}(\omega)$ and $\hat{P}_{\text{out},n}(\omega)$ are the spectral power estimates of the reference microphone observation and the output, respectively, when the interferences are active alone and C_d is a constant to normalize the desired source's gain. The performance is also given in terms of the source distortion measure (SD), defined as

$$\text{SD} = 10\log_{10}\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\left(\frac{1}{C_d}\right)\hat{P}_{\text{in},s}(\omega) - \hat{P}_{\text{out},s}(\omega)\right| d\omega\right), \tag{36}$$

where $\hat{P}_{\text{in},s}(\omega)$ and $\hat{P}_{\text{out},s}(\omega)$ are the spectral power estimates of the reference microphone observation and the output, respectively, when the desired source is active alone. The source distortion is the mean output spectral power deviation from the observed single sensor spectral power. Ideally, the distortion is zero.

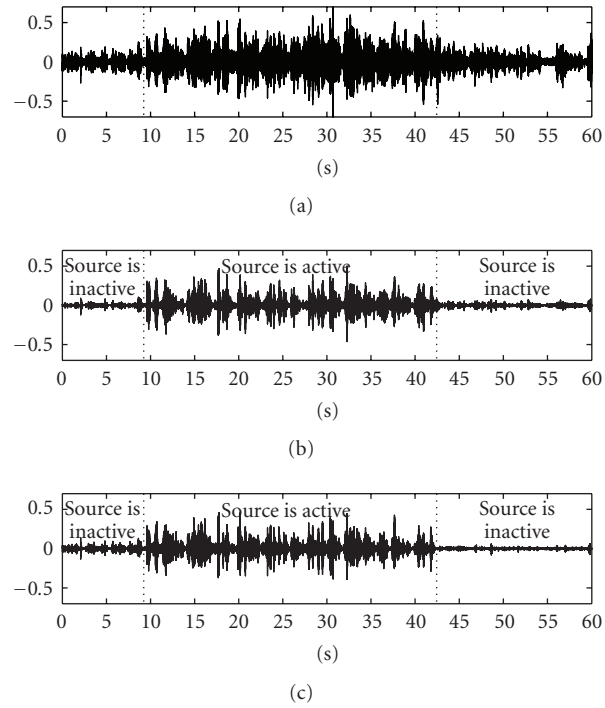

Figure 6: Source 1 is the desired source: time domain plots of outputs from (a) the fixed multiobjective optimal beamformer, (b) the postfilter with power spectral estimation using inner product and (c) the postfilter with power spectral estimation using joint diagonalization.

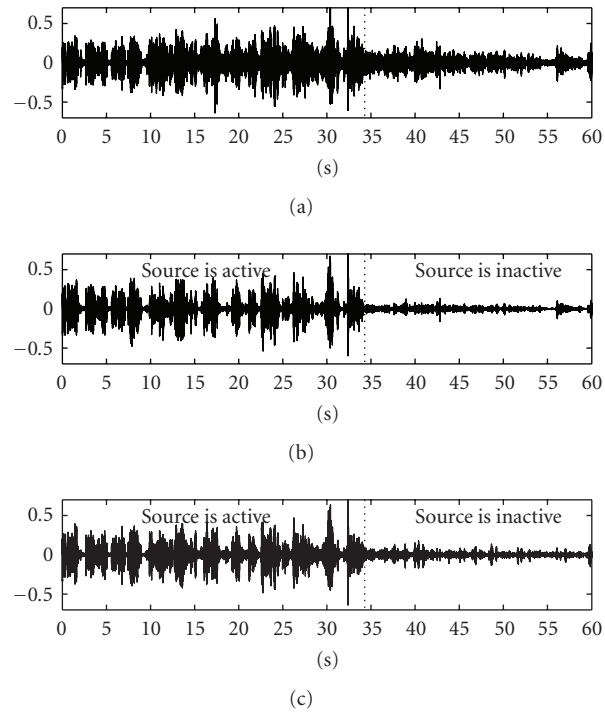

Figure 7: Source 2 is the desired source: time domain plots of outputs from (a) the fixed multiobjective optimal beamformer, (b) the postfilter with power spectral estimation using inner product, and (c) the postfilter with power spectral estimation using joint diagonalization.

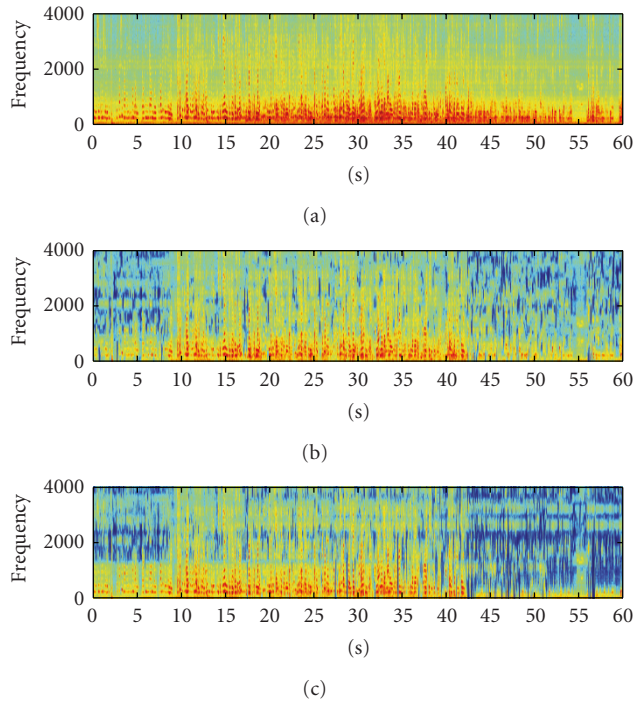

Figure 8: Source 1 is the desired source: spectrograms of outputs from (a) the fixed multiobjective optimal beamformer, (b) the postfilter with power spectral estimation using an inner product, and (c) the postfilter with power spectral estimation using joint diagonalization.

Here, one speaker is viewed as the desired signal while others are undesired or interference signals. Obviously, the suppression levels for each undesired source are different depending on the spatial differences between its location and the location of the desired source. However, we consider all the undesired signals as one interference signal for evaluating the IS level for the proposed methods.

The proposed beamformers are employed to enhance a desired speech signal. Figures 6 and 8 show, respectively, the time domain and the spectrogram plots of (a) the fixed multiobjective optimal beamformer, (b) the postfilter with PSD estimation using an inner product, and (c) the postfilter with PSD estimation using joint diagonalization, with the desired source chosen as the 1st source. Also, the time domain and the spectrogram plots of the output for the 2nd source are illustrated in Figures 7 and 9, respectively.

As the suppression and distortion levels are different for the active and inactive periods of the desired source, these two cases are analyzed separately.

6.1. Active time of the desired source

Evaluations are obtained for the periods in which the desired source is active. For example, the periods [9 seconds, 42 seconds] and [0 second,34 seconds] are considered as the active time for the 1st and the 2nd sources, respectively. Also, Figures 6 and 7 show the active time for the corresponding desired sources. The active periods are viewed as "source is active."

The desired source is chosen as one of the four speech signals. Table 1 shows the IS and the SD levels in the output of the delay and sum beamformer, the multiobjective optimal beamformer, the postfilter with PSD estimation using an inner product, and the postfilter with PSD estimation using joint diagonalization. The delay and sum beamformer forms a beam towards a specified direction by matching the delay such that signals from that direction will be reinforced (summed together with matching delay).

The IS level for the delay and sum beamformer ranges from 0.3 to 1.3 dB depending on the desired source position. The IS level for the multiobjective optimal beamformer ranges from 5 to 6.57 dB depending on the desired source position. The results show that the multiobjective optimal beamformer achieves a significant improvement in the IS levels over the delay and sum beamformer. The postfilters improve further the IS levels of the multiobjective optimal beamfomer outputs. More specifically, the postfilter with PSD estimation using an inner product improves approximately 3 dB in IS level over the fixed multiobjective optimal beamformer for all the desired sources. The postfilter with PSD estimation using joint diagonalization improves approximately 2.5 dB in IS level for all the desired sources.

The speakers 1 and 4 have slightly better IS than the other two speakers. This is due to the fact that those speakers' positions are more spatially separated when compared to the other positions. From simulation results, the postfilter with PSD estimation using inner product has a slightly higher IS level than the one using joint diagonalization. On the other hand, the postfilter using joint diagonalization has a slightly lower SD than the one with inner product. In general, all the outputs have low SD levels, leading to low distortion of the desired source.

6.2. Inactive time of the desired source

Evaluations are also obtained for the periods in which the desired source is inactive. For example, the time periods [0 second, 9 seconds] and [42 seconds, 60 seconds] are inactive periods for the 1st source, (see Figure 6). Thus, evaluation is performed for the combining outputs of both periods. Also, the time period [34 seconds, 60 seconds] is inactive period for the 2nd source (see Figure 7). In Figures 6 and 7, the inactive periods for the corresponding desired sources are viewed as "source is inactive." In addition, the signal to interference ratio (SIR) is zero in the inactive source periods and there is only an IS measure for the evaluation.

Table 2 shows the IS levels for the outputs of the delay and sum beamformer, the fixed multiobjective optimal beamformer, the postfilter with PSD estimation using an inner product and the postfilter with PSD estimation using joint diagonalization. The range of IS levels for the delay and sum beamformer and the fixed multiobjective optimal beamformer remains approximately the same as for the source active periods. The IS of the postfilters with PSD estimations, however, is significantly improved over the previous case

Table 1: Desired source is active: IS and SD levels of delay and sum beamformer (DLSB) output, fixed multiobjective optimal beamformer (FMOB) output, the postfilter with power spectral estimation using an inner product (PF & IPT), and the postfilter with power spectral estimation using joint diagonalization (PF & JDG).

Desired source	DLSB		FMOB		PF & IPT		PF & JDG	
	IS dB	SD dB	IS dB	SD dB	IS dB	SD dB	IS dB	SD dB
1	1.3	−37.4	6.8	−29.2	9.5	−27.9	9.2	−28.2
2	0.3	−35.8	5.7	−26.6	9.1	−25.4	8.0	−26.0
3	0.7	−37.4	5.0	−28.2	7.9	−26.3	7.1	−26.9
4	0.8	−37	6.3	−26	8.9	−25.0	8.6	−25.5

Table 2: Desired source is nonactive: IS levels for the outputs of delay and sum beamformer (DLSB), the fixed multiobjective optimal beamformer (FMOB), postfilter with power spectral estimation using inner product (PF & IPT), and postfilter with power spectral estimation using joint diagonalization (PF & JDG).

Desired source	DLSB IS (dB)	FMOB IS (dB)	PF & IPT IS (dB)	PF & JDG IS (dB)
1	1.3	7.2	17.7	17.1
2	0.3	6.7	17.2	16.3
3	0.7	5.4	15.9	14.5
4	0.8	6.8	17.1	16.9

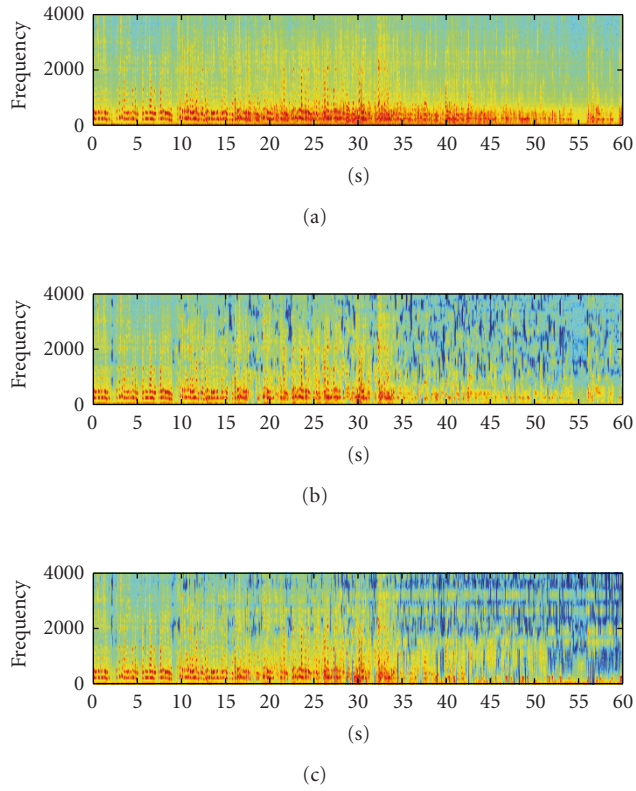

Figure 9: Source 2 is the desired source: spectrograms of outputs from (a) the fixed multiobjective optimal beamformer, (b) the postfilter with power spectral estimation using an inner product, and (c) the postfilter with power spectral estimation using joint diagonalization.

where the desired source is active. More specifically, the postfilter using an inner product improves approximately 10 dB over the fixed multiobjective optimal beamformer output for all desired sources. Similarly, the postfilter with PSD estimation using joint diagonalization improves approximately 9 dB for all the desired sources.

Similar to the case where the desired source is active, better IS levels are obtained for the 1st and the 4th speakers. Also, the postfilter with PSD estimation using an inner product has a slightly higher suppression level than the one using joint diagonalization.

From the simulation results, the postfilter with spectral estimation using an inner product has a slightly higher interference suppression level than the postfilter with spectral estimation using the joint diagonalization. This also comes with a higher computational complexity as the number of real multiplications required for each frequency bin by the first estimation method is higher than the second method, for example, $4(2L^2+1)$ versus $4L^2$, (see Sections 5.1 and 5.2). A limitation of the proposed methods is that calibration is required for the spatial correlation matrix estimation. Further work is required to investigate the near-field estimation models of the spatial correlation matrix with on-time spatial information update.

7. CONCLUSIONS

In this paper, a two-stage beamformer structure is proposed for speech enhancement in a multispeaker environment. In the first stage, a fixed multiobjective optimal beamformer is designed to spatially extract the desired source. In the second stage, a postfilter technique is used to further enhance the extraction process. Two different multichannel power spectral estimation methods have been proposed and evaluated. Both methods are capable of estimating the desired source PSD in a multispeaker environment. Evaluations in a real environment show that both methods have similar suppression capability and comparable distortion levels. The postfilter with spectral estimation using inner product has a slightly higher suppression level than the method using joint diagonalization with a higher computational complexity.

ACKNOWLEDGMENTS

WATRI is a joint venture between the University of Western Australia and Curtin University of Technology. This work

was sponsored by National ICT Australia (NICTA). NICTA is funded through the Australian Government's Backing Australia's Ability initiative, in part through Australian Research Council.

REFERENCES

[1] J. Benesty, S. Makino, and J. Chen, Eds., *Speech Enhancement*, Springer, Berlin, Germany, 2005.

[2] M. Brandstein and D. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*, Springer, Berlin, Germany, 2001.

[3] S. Nordebo, I. Claesson, and S. Nordholm, "Adaptive beamforming: spatial filter designed blocking matrix," *IEEE Journal of Oceanic Engineering*, vol. 19, no. 4, pp. 583–590, 1994.

[4] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Transactions on Signal Processing*, vol. 50, no. 9, pp. 2230–2244, 2002.

[5] N. Grbić, S. Nordholm, and A. Cantoni, "Optimal FIR subband beamforming for speech enhancement in multipath environments," *IEEE Signal Processing Letters*, vol. 10, no. 11, pp. 335–338, 2003.

[6] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, Upper Saddle River, NJ, USA, 4th edition, 2001.

[7] H. Q. Dam, S. Y. Low, S. Nordholm, and H. H. Dam, "Adaptive microphone array with noise statistics updates," in *Proceedings of the International Symposium on Circuits and Systems (ISCAS '04)*, vol. 3, pp. 433–436, Vancouver, British Columbia, Canada, May 2004.

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

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

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

[11] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP '88)*, vol. 5, pp. 2578–2581, New York, NY, USA, April 1988.

[12] I. A. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 709–716, 2003.

[13] I. Cohen, "Multichannel post-filtering in nonstationary noise environments," *IEEE Transactions on Signal Processing*, vol. 52, no. 5, pp. 1149–1160, 2004.

[14] Y. Huang, J. Benesty, and J. Chen, "Separation and dereverberation of speech signals with multiple microphones," in *Speech Enhancement*, chapter 12, pp. 271–298, Springer, Berlin, Germany, 2005.

[15] Y. Cao, S. Sridharan, and M. Moody, "Speech enhancement iby simulation of cocktail party effect with neural network controlled iterative filter," in *Proceedings of the 4th International Symposium on Signal Processing and Its Applications (ISSPA '96)*, vol. 2, pp. 541–544, Gold Cost, Australia, August 1996.

[16] J. M. de Haan, N. Grbić, I. Claesson, and S. Nordholm, "Design of oversampled uniform DFT filter banks with delay specification using quadratic optimization," in *Proceedings of the IEEE Interntional Conference on Acoustics, Speech, and Signal Processing (ICASSP '01)*, vol. 6, pp. 3633–3636, Salt Lake, Utah, USA, May 2001.

[17] H. Q. Dam, S. Nordholm, H. H. Dam, and S. Y. Low, "Maximum likelihood estimation and Cramer-Rao lower bounds for the multichannel spectral evaluation in hands-free communication," in *Proceedings of Asia-Pacific Conference on Communications (APCC '05)*, pp. 961–964, Perth, Australia, October 2005.

[18] H. Q. Dam, S. Nordholm, H. H. Dam, and S. Y. Low, "Post-filtering with multichannel power spectral estimation using joint diagonalization in multi-speaker environments," in *Proceedings of Asia-Pacific Conference on Communications (APCC '06)*, pp. 1–5, Busan, Korea, August 2006.

[19] S. Nordholm, I. Claesson, and M. Dahl, "Adaptive microphone array employing calibration signals: an analytical evaluation," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 241–252, 1999.

[20] G. L. Fudge and D. A. Linebarger, "Calibrated generalized sidelobe canceller for wideband beamforming," *IEEE Transactions on Signal Processing*, vol. 42, no. 10, pp. 2871–2875, 1994.

[21] J. M. Sachar, H. F. Silverman, and W. R. Patterson III, "Position calibration of large-aperture microphone arrays," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '02)*, vol. 2, pp. 1797–1800, Orlando, Fla, USA, May 2002.

[22] G. Strang, *Linear Algebra and Its Applications*, Academic Press, New York, NY, USA, 1976.

[23] J.-F. Cardoso and A. Souloumiac, "Jacobi angles for simultaneous diagonalization," *SIAM Journal on Matrix Analysis and Applications*, vol. 17, no. 1, pp. 161–164, 1996.