The 2010 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 18-22, 2010, Taipei, Taiwan

Sound Source Separation by Using Matched Beamforming and Time-Frequency Masking

Jounghoon Beh[*], Taekjin Lee[†], David Han[‡], Hanseok Ko[◊]
[*]*UMIACS, University of Maryland, College Park, USA*
[†]*UBICOD Co. Ltd, KOREA*
[‡]*Office of Naval Research, USA*
[◊]*Korea University, KOREA*
jhbeh@umiacs.umd.edu, hsko@korea.ac.kr

***Abstract*— This paper proposes a two-stage algorithm to separate two sound sources by using matched beamforming and time-frequency masking techniques. At first, beamforming was used to separate the sound mixtures back to the original sources while preserving the original contents to the maximum extent. The residual interference was then suppressed by the time-frequency masking technique. A sequential least squares method was used in developing a matched beamformer to estimate the relative transfer function (RTF). From experimental results, it has been shown that the proposed method exhibits improved performance in sound source separation compared to conventional methods. Signal enhanced factor (SEF) was improved by an average of 8.39 dB over the baseline.**

I. Introduction

The key objective of a speech based human-robot interface is to accurately recognize operator's acoustic commands. In real life environments, the operator's voice command is often corrupted by noise such as competing voices as in the case of cocktail party problem or nonverbal background noise that would drastically degrade performance of any speech recognition system.

To alleviate this, several approaches were suggested on the subject of separating the mixed sounds into the respective sources. Independent component analysis (ICA) and beamforming are some of the representative methods. ICA based method utilizes independent characteristics of each sound source by assuming that each input source is statistically independent. In contrast to ICA, beamforming based method needs to know the direction of the desired source to reduce other speech and noise. The method separates the audible sources by separating them in terms of their directions. Time-frequency masking (TFM) method [1] separates sound sources by masking unwanted sounds in the time-frequency domain. The method primarily relies on clustering of the mixed signals with respect to their amplitudes and time delays.

The motivation of this paper is to develop a sound source separation system for the purpose of enhancing the quality of the robot operator's speech. For the separation of the sound sources, a matched beamformer is employed first with an assumption that the locations of the operator and other competing speakers are known as assumed in [2]. The matched beamformer employs a least square method to estimate the impulse response, also known as a transfer function (TF), between the sound sources and the input signals of the microphone array. Existing least square based methods require a set of input data blocks for initial calibration [3]. However, this approach may result poor performance due to problems such as memory insufficiency or low performance of the TF estimation in the case of moving sound sources. This would affect the sound source separation process which would eventually degrade the overall performance. An adaptive form of a least squares solution using the least mean squares (LMS) was introduced in previous studies [4]. However, this led to a significant increase in computational load. To counter these problems and to further improve the performance, this paper suggests the use of a sequential least square method.

The remainder of this paper is organized as follows. Section II describes the proposed algorithm in detail. Experimental results are discussed in Section III. Concluding remarks are presented in Section IV.

II. Proposed Algorithm

Let $s(m)$ and $c(m)$ denote respectively the voice signal of operator and competing speaker where m is the discrete time index. Then the observed signal at i^{th} microphone of array, $y_i(m)$, is assumed to be given by

$$y_i(m) = h_i(m) * s(m) + g_i(m) * c(m) + n_i(m), i = 1,..., M \quad (1)$$

where $h_i(m)$ and $g_i(m)$ represent impulse responses between the i^{th} microphone and each of the two sources. M denotes the number of microphones consisting of the array, and $n_i(m)$ is ambient noise signal which is assumed to be stationary and uncorrelated with the source signal.

The signal model of interest in this paper is formed in the Short Time Fourier Transform (STFT) domain. $y_i(m)$ is segmented into overlapping frames with analysis window of size W and the shift size of Z. We form the signal in the l^{th} frame as

$$y_i(l, v) = y_i(lZ + v)w(v), \quad 0 \le v \le W - 1, \quad (2)$$

where $w(v)$ represents an analysis window. By using Fourier transform, (2) becomes

$$Y_i(l, \omega) = \sum_{v=0}^{W-1} y_i(l, v) e^{-j\omega v} \quad (3)$$

In STFT domain, (1) can be rewritten as

$$Y_i(l, \omega) = H_i(\omega) S(l, \omega) + G_i(\omega) C(l, \omega) + N_i(l, \omega) \quad (4)$$

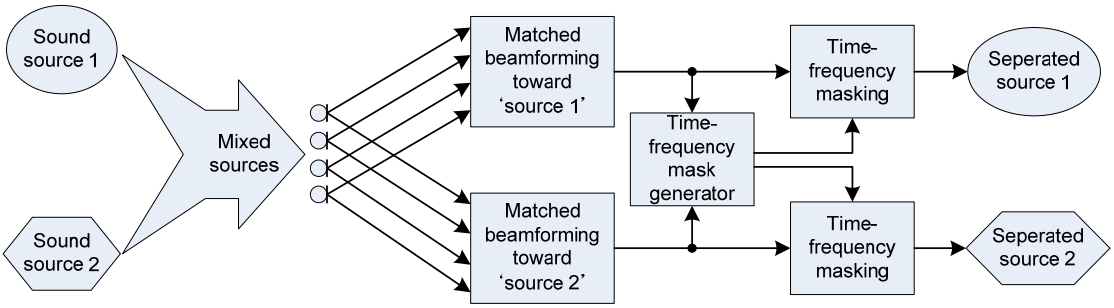

Fig. 1. Block diagram of the proposed algorithm.

Our goal in this paper is to recover original signals $S(l,\omega)$ and $C(l,\omega)$ from the given observation $Y_i(l,\omega)$ by using the matched beamformer and TFM. We illustrate the proposed algorithm in terms of a block diagram as shown in Fig. 1. First, the microphone array determines if there was any utterance from any of the parties by measuring the acoustic energy level collected and proceeds to estimating the associated TF once the collected level exceeds a threshold. We then proceed to estimate relative TFs (RFTs) that represent the relationships between the reference microphone and other microphones of the array. This is necessary as a part of a calibration process prior to the matched beamforming. Using the estimated RTFs from the prior step, matched beamformers are developed, with beam width being constrained to the degree such that preservation of the original source contents was maximized. The resultant outputs from the matched beamfomer are roughly separated back to the two person's voice, however each of the separated signals may yet contain the acoustic contents from the other party's voice. To further reduce the residual voice contaminated from the other speaker, we employ the TFM based on the spectral power ratios between the two separated signals obtained from the matched beamforming.

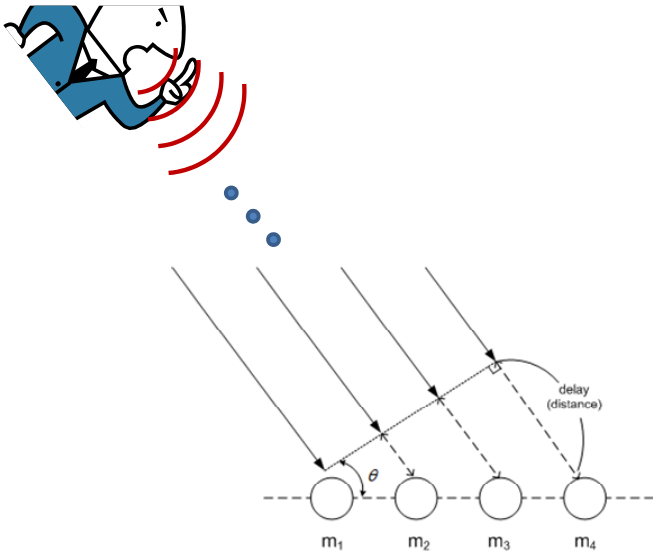

Fig. 2. The incidence of acoustic wave to microphone array in planar manner.

A. *Detecting sound from a direction of interest*

For estimating RFT corresponding to a direction of interest, it is necessary to obtain certain amount of acoustic data from the direction. In order to do so, it needs to be determined whether an acoustic event occurred or not. Occurrence of an acoustic event is determined by examining its acoustic energy intensity from its direction. Direction of the speaker is determined from a visual sensor as it was assumed in [2]. As the operator gives verbal commands to the robot, the audible signal reaches the array with an incidence angle θ. In our implementation we used 4-channel microphone array, denoted by $m_1 \sim m_4$ as shown in Fig. 2, and we employed the far-field assumption that the propagated wave reaches the array in planar manner. With respect to the incidence angle θ, signal arrives at each microphone with corresponding time-delays.

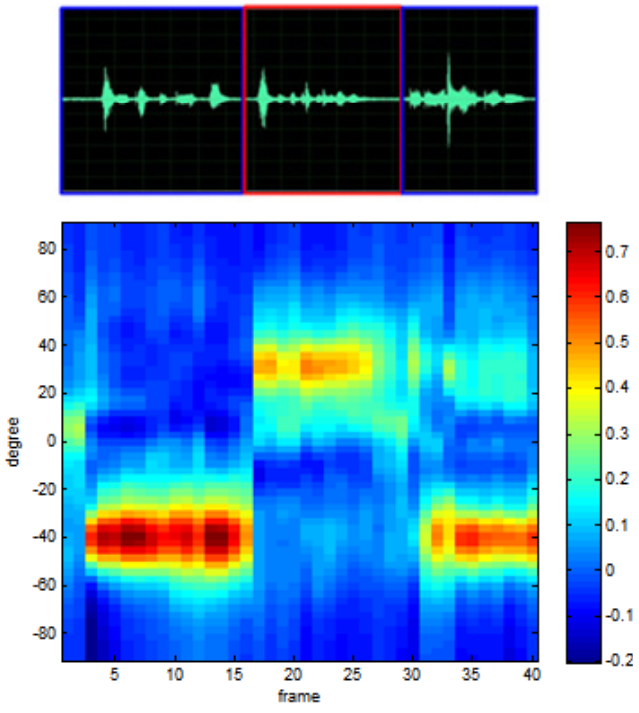

Fig. 3. Energy intensity as a function of time and angle of the sound source location.

Steered response power phase transform (SRP-PHAT) [5] is used for examining the acoustic energy intensity. SRP-PHAT value, $SP(l, \theta)$, as a function of frame l and

direction θ is obtained by

$$SP(l,\theta)=\sum_{i=1}^{M}\sum_{j=1}^{M}\int_{0}^{2\pi}\Psi_{ij}(l,\omega)Y_i(l,\omega)Y_j^*(l,\omega)e^{j\omega\tau_{ij}(\theta)}d\omega \quad (5)$$

where $\tau_{ij}(\theta)$ denotes the time-delay of the signal between the i^{th} and j^{th} microphones with respect to θ. The phase transform (PHAT) weighting function $\Psi_{ij}(l,\omega)$ is defined as

$$\Psi_{ij}(l,\omega)\equiv\frac{1}{|Y_i(l,\omega)Y_j^*(l,\omega)|}. \quad (6)$$

Fig. 3 illustrates the energy intensity by using (5) when two speakers alternatively utter, and they are located at angles of -40° and 30° respectively. After examining the energy intensity, we determine whether an acoustic event occurred in the direction of interest, e.g. operator's angle.

B. RTF estimation

RTFs of $H_i(\omega)$ and $G_i(\omega)$ in (4) are defined as ratios of TFs between the i^{th} microphone and the reference one. Note that we choose the leftmost one, which is the 1st microphone, as a reference.

$$P_i(\omega)\equiv\frac{H_i(\omega)}{H_1(\omega)},\quad i=1,...,M, \quad (7a)$$

$$Q_i(\omega)\equiv\frac{G_i(\omega)}{G_1(\omega)},\quad i=1,...,M. \quad (7b)$$

We propose a RTF estimation method which employs the sequential-mode least squares (SLS) [6]. There are two advantages of the proposed method over the conventional methods [3] employing the batch-mode least squares (BLS). First, it provides flexibility in collecting the acoustic data to estimate the RFTs. For the methods utilizing the BLS, it is required that acoustic data only from corresponding direction should be collected for a set period of time. In our experiments of implementing the BLS based method, at least 3.2 seconds of collection period was needed for sufficient RTF estimations to guarantee reasonable performance of the matched beamformer. Our proposed method accomplishes the estimation by updating the RFT on frame-by-frame basis, therefore there is no threshold acoustic collection period needed.

The other advantage is memory efficiency achievable in hardware based processing. In the BLS implementation, in terms of bytes, the memory required is {(# of microphones – 1) × (2 × # of frequency bins) × (# of frame to be used in estimation) × (# of bytes w.r.t. variable type)} per a source. As we mentioned in this section, RTF is estimated per frame. It does not need to arrange memory of a hardware for the estimation.

An RTF is estimated when the current input frame is determined to contain an acoustic event caused by a person (operator or competing speaker). Estimation of an RTF associated with an operator location is considered as an example. Observation $Y_i(l,\omega)$ in (4) is rearranged by using (7a) as

$$\begin{aligned}Y_i(l,\omega)&=H_i(\omega)S(l,\omega)+N_i(l,\omega)\\&=P_i(\omega)Y_1(l,\omega)+U_i(l,\omega),\end{aligned} \quad (8)$$

where $U_i(l,\omega)$ is formed as

$$U_i(l,\omega)=N_i(l,\omega)-P_i(\omega)N_1(l,\omega). \quad (9)$$

In (8) and (9), we assume that the analysis interval of the STFT is sufficiently long enough for the observed signal in l^{th} frame, $y_i(l,r)$, to be considered stationary. In addition, we have assumed that the ambient noise $n_i(m)$ is stationary. Thus, the cross power spectral density (CPSD) between Y_i and Y_1 can be written as

$$\Phi_{Y_iY1}(l,\omega)=P_i(\omega)\Phi_{Y1Y1}(l,\omega)+\Phi_{U_iY1}(\omega). \quad (10)$$

Note that since U_i is uncorrelated with S_i, Φ_{U_iY1} is independent of l. We estimate $\Phi_{Y_iY1}(l,\omega)$ by the recursive averaging along the frame. Let $\hat{\Phi}_{Y_iY1}(l,\omega)$ denote estimates of $\Phi_{Y_iY1}(l,\omega)$. Then, (10) is rewritten as

$$\begin{aligned}\hat{\Phi}_{Y_iY1}(l,\omega)&=P_i(\omega)\hat{\Phi}_{Y1Y1}(l,\omega)+\hat{\Phi}_{U_iY1}(l,\omega)\\&=P_i(\omega)\hat{\Phi}_{Y_iY1}(l,\omega)+\Phi_{U_iY1}(\omega)+\varepsilon_i(l,\omega),\end{aligned} \quad (11)$$

where $\varepsilon_i(l,\omega)=\Phi_{U_iY1}(l,\omega)-\hat{\Phi}_{U_iY1}(l,\omega)$ is the estimation error. We can now consider the acoustic data in a certain duration corresponding to K frames for estimating $P_i(\omega)$. Via the BLS approach, $P_i(\omega)$ can be obtained from the following equations [3][7]:

$$\begin{bmatrix}\hat{\Phi}_{Y_iY1}^{(1)}(\omega)\\ \hat{\Phi}_{Y_iY1}^{(2)}(\omega)\\ \vdots\\ \hat{\Phi}_{Y_iY1}^{(K-1)}(\omega)\\ \hat{\Phi}_{Y_iY1}^{(K)}(\omega)\end{bmatrix}=\begin{bmatrix}\hat{\Phi}_{Y1Y1}^{(1)}(\omega) & 1\\ \hat{\Phi}_{Y1Y1}^{(2)}(\omega) & 1\\ \vdots & \\ \hat{\Phi}_{Y1Y1}^{(K-1)}(\omega) & 1\\ \hat{\Phi}_{Y1Y1}^{(K)}(\omega) & 1\end{bmatrix}\begin{bmatrix}P_i(\omega)\\ \Phi_{U_iY1}(\omega)\end{bmatrix}+\begin{bmatrix}\varepsilon_i^{(1)}(\omega)\\ \varepsilon_i^{(2)}(\omega)\\ \vdots\\ \varepsilon_i^{(K-1)}(\omega)\\ \varepsilon_i^{(K)}(\omega)\end{bmatrix} \quad (12)$$

where superscript $^{(\cdot)}$ denote the frame number of data used for the BLS.

The idea behind the SLS is to recursively update the least squares estimate as new observations are acquired [6]. The following vectors are defined for sequentially solving the equation (12).

$$\mathbf{y}_i^{(k)}=[\hat{\Phi}_{Y_iY1}^{(1)}(\omega)\quad\hat{\Phi}_{Y_iY1}^{(2)}(\omega)\quad ...\hat{\Phi}_{Y_iY1}^{(k)}(\omega)]^T, \quad (13)$$

$$\mathbf{a}_k=[\hat{\Phi}_{Y1Y1}^{(k)}(\omega)\quad 1], \quad (14)$$

$$\mathbf{A}_k=\begin{bmatrix}\hat{\Phi}_{Y1Y1}^{(1)}(\omega) & 1\\ \hat{\Phi}_{Y1Y1}^{(2)}(\omega) & 1\\ \vdots & \\ \hat{\Phi}_{Y1Y1}^{(k)}(\omega) & 1\end{bmatrix}=\begin{bmatrix}\mathbf{A}_{k-1}\\ \hline \mathbf{a}_k\end{bmatrix}, \quad (15)$$

$$\mathbf{n}_i^{(k)}=[\varepsilon_i^{(1)}(\omega)\quad\varepsilon_i^{(2)}(\omega)\quad ...\varepsilon_i^{(k)}(\omega)]^T. \quad (16)$$

$$\boldsymbol{\theta}_i=[R_i(\omega)\quad\Phi_{U_iY1}(\omega)]^T \quad (17)$$

Then, (12) can be rewritten as

$$\mathbf{y}_i^{(k)}=\begin{bmatrix}\mathbf{y}_i^{(k-1)}\\ \hline \hat{\Phi}_{Y_iY1}^{(k)}(\omega)\end{bmatrix}, \quad (18a)$$

where

$$\mathbf{y}_i^{(k-1)}=\mathbf{A}_{k-1}\boldsymbol{\theta}_i+\mathbf{n}_i^{(k-1)}, \quad (18b)$$

$$\hat{\Phi}_{Y_iY1}^{(k)}(\omega)=\mathbf{a}_k\boldsymbol{\theta}_i+\varepsilon_i^{(k)}(\omega). \quad (18c)$$

Let $\hat{\boldsymbol{\theta}}_i^{(k)}$ denote the SLS solution given the measurement $\hat{\boldsymbol{\Phi}}_{YiY1}^{(k)}(\omega)$. By using (14) and (18), and the given measurement vector $\mathbf{y}_i^{(k)}$, $\hat{\boldsymbol{\theta}}_i^{(k)}$ is determined as follow:

$$\begin{aligned}\hat{\boldsymbol{\theta}}_i^{(k)} &= (\mathbf{A}_k^H\mathbf{A}_k)^{-1}\mathbf{A}_k^H\mathbf{y}_i^{(k)} \\ &= (\mathbf{A}_{k-1}^H\mathbf{A}_{k-1}+\mathbf{a}_k^H\mathbf{a}_k)^{-1}(\mathbf{A}_{k-1}^H\mathbf{y}_i^{(k-1)}+\mathbf{a}_k^H\hat{\boldsymbol{\Phi}}_{YiY1}^{(k)}(\omega)).\end{aligned} \tag{19}$$

Let $\mathbf{D}_k$ denote the inverse of the Gram matrix of $\mathbf{A}_k$ such that

$$\mathbf{D}_k = (\mathbf{A}_k^H\mathbf{A}_k)^{-1}. \tag{20}$$

Use of the matrix-inversion lemma and (20) leads us to

$$\begin{aligned}(\mathbf{A}_{k-1}^H\mathbf{A}_{k-1}+\mathbf{a}_k^H\mathbf{a}_k)^{-1} &= (\mathbf{D}_{k-1}^{-1}+\mathbf{a}_k^H\mathbf{a}_k)^{-1} \\ &= \mathbf{D}_{k-1}-\mu_k\mathbf{D}_{k-1}\mathbf{a}_k^H\mathbf{a}_k\mathbf{D}_{k-1}\end{aligned}. \tag{21a}$$

where

$$\mu_k^{-1} = 1+\mathbf{a}_k\mathbf{D}_{k-1}\mathbf{a}_k^H. \tag{21b}$$

Then, substituting (20) and (21) into (19) yields the SLS solution which is recursively updated.

$$\begin{aligned}\hat{\boldsymbol{\theta}}_i^{(k)} &= (\mathbf{D}_{k-1}-\mu_k\mathbf{D}_{k-1}\mathbf{a}_k^H\mathbf{a}_k\mathbf{D}_{k-1})\{\mathbf{A}_{k-1}^H\mathbf{y}_i^{(k-1)}+\mathbf{a}_k^H\hat{\boldsymbol{\Phi}}_{YiY1}^{(k)}(\omega)\} \\ &= \mathbf{D}_{k-1}\mathbf{A}_{k-1}^H\mathbf{y}_{k-1}-\mu_k\mathbf{D}_{k-1}\mathbf{a}_k^H\mathbf{a}_k\mathbf{D}_{k-1}\mathbf{A}_{k-1}^H\mathbf{y}_i^{(k-1)} \\ &\quad+\mathbf{D}_{k-1}\mathbf{a}_k\hat{\boldsymbol{\Phi}}_{YiY1}^{(k)}(\omega)-\mathbf{D}_{k-1}\mathbf{a}_k^H(1-\mu_k)\hat{\boldsymbol{\Phi}}_{YiY1}^{(k)}(\omega) \\ &= \hat{\boldsymbol{\theta}}_i^{(k-1)}+\mu_k\mathbf{D}_{k-1}\mathbf{a}_k^H\{\hat{\boldsymbol{\Phi}}_{YiY1}^{(k)}(\omega)-\mathbf{a}_k\hat{\boldsymbol{\theta}}_i^{(k-1)}\}\end{aligned} \tag{22}$$

$P_i(\omega)$, the first element of $\hat{\boldsymbol{\theta}}_i^{(k)}$ which estimates the RTF between the i^{th} microphone and the speaker is the key signal to be captured here. The other RTF, $Q_i(\omega)$, can be obtained via the similar procedure. Since we have assumed the background noise is stationary, by using (11) and (18c) the difference between the estimated CPSD $\hat{\boldsymbol{\Phi}}_{YiY1}^{(k)}(\omega)$ and $\mathbf{a}_k\hat{\boldsymbol{\theta}}_i^{(k-1)}$ would lead to the estimation error of RFT as follow:

$$\begin{aligned}\hat{\boldsymbol{\Phi}}_{YiY1}^{(k)}(\omega)-\mathbf{a}_k\hat{\boldsymbol{\theta}}_i^{(k-1)} &= P_i(\omega)\hat{\boldsymbol{\Phi}}_{Y1Y1}^{(k)}+\hat{\boldsymbol{\Phi}}_{UiY1}(\omega)-\hat{P}_i^{(k-1)}(\omega)\hat{\boldsymbol{\Phi}}_{Y1Y1}^{(k)}(l,\omega) \\ &\quad-\hat{\boldsymbol{\Phi}}_{UiY1}^{(k-1)}(\omega) \\ &\approx \hat{\boldsymbol{\Phi}}_{Y1Y1}^{(k)}(P_i(\omega)-\hat{P}_i^{(k-1)}(\omega))\end{aligned} \tag{23}$$

Equation (22) and (23) tells us that the error is reflected to the update of the current estimate of RTF with the gain $\mu_k\mathbf{D}_{k-1}\mathbf{a}_k^H$.

C. Matched beamformer

The matched beamformer was originally proposed as a part of *dual transfer function generalized sidelobe canceller* [7]. In the proposed algorithm, its role is to form the beam toward a person while nulling the audio signal coming from the other person's location. We briefly describe the procedure in this section.

Let $\mathbf{P}(\omega)$ and $\mathbf{Q}(\omega)$ denote the RTF vectors formed by using (7) such that

$$\mathbf{P}(\omega) = [1 \quad P_2(\omega) \quad \ldots \quad P_M(\omega)]^T, \tag{24a}$$

$$\mathbf{Q}(\omega) = [1 \quad Q_2(\omega) \quad \ldots \quad Q_M(\omega)]^T. \tag{24b}$$

With estimates of the RTFs, the matched beamformer gain for forming beam toward the operator is given by

$$\mathbf{W}_S(\omega) = \frac{\dfrac{\hat{\mathbf{P}}(\omega)}{\|\hat{\mathbf{P}}(\omega)\|^2}-\rho(\omega)\dfrac{\hat{\mathbf{Q}}(\omega)}{\|\hat{\mathbf{P}}(\omega)\|\|\hat{\mathbf{Q}}(\omega)\|}}{1-\|\rho(\omega)\|^2}F_S(\omega) \tag{25}$$

where $\rho(\omega)$ is defined by

$$\rho(\omega) \equiv \frac{\hat{\mathbf{Q}}^H(\omega)\hat{\mathbf{P}}(\omega)}{\|\hat{\mathbf{Q}}^H(\omega)\|\,\|\hat{\mathbf{P}}(\omega)\|} \tag{26}$$

with the cosine of the angle between the $\mathbf{Q}(\omega)$ and $\mathbf{P}(\omega)$ forming in an inner product space. $F_S(\omega)$ is the desired filter response of the matched beamformer with respect to the source $S(l,\omega)$ in (4). Likewise, the matched beamformer gain toward competing speaker is given by

$$\mathbf{W}_C(\omega) = \frac{\dfrac{\hat{\mathbf{Q}}(\omega)}{\|\hat{\mathbf{Q}}(\omega)\|^2}-\rho^*(e^{j\omega})\dfrac{\hat{\mathbf{P}}(\omega)}{\|\hat{\mathbf{Q}}(\omega)\|\|\hat{\mathbf{P}}(\omega)\|}}{1-\|\rho(\omega)\|^2}F_C(\omega). \tag{27}$$

where $F_C(\omega)$ is the desired filter response of the matched beamformer with respect to the source $C(l,\omega)$.

Let $\mathbf{Y}(l,\omega)$ denote the input signal vector of array, i.e. $[Y_1(l,\omega) \;\; Y_2(l,\omega) \;\; \ldots \;\; Y_M(l,\omega)]^T$, and let $\widetilde{S}(l,\omega)$ and $\widetilde{C}(l,\omega)$ denote the beamformer outputs with respect to the operator and a competing speaker, respectively. By using (25) and (27), the beamformer outputs are written as

$$\begin{bmatrix}\widetilde{S}(l,\omega)\\ \widetilde{C}(l,\omega)\end{bmatrix} = \begin{bmatrix}\mathbf{W}_S^H(\omega)\\ \mathbf{W}_C^H(\omega)\end{bmatrix}\mathbf{Y}(l,\omega). \tag{28}$$

D. Time-frequency masking

Due to imperfect estimations of the RFTs, the beamformer outputs may yet contain acoustic contents from the other party's voice. We employ the spectral power based TFM to further reduce the noise. The TFM is based on the windowing-disjoint orthogonal (WDO) [1] assumption which means that two signals in time-frequency domain such as the STFT domain are assumed not to coexist at the same time and same frequency bin. In application to our problem, it can be stated concisely as

$$S(l,\omega)C(l,\omega) = 0, \quad \forall l,\omega. \tag{29}$$

Since the beamformer gain function is a linear operator, the WDO assumption can be applied to the beamformer output as

$$\widetilde{S}(l,\omega)\widetilde{C}(l,\omega) = 0, \quad \forall l,\omega. \tag{30}$$

We propose a mask function based on the spectral power ratio which is defined by

$$R_S(l,\omega) = 10\log_0\left(\frac{\|\widetilde{S}(l,\omega)\|^2}{\|\widetilde{C}(l,\omega)\|^2}\right). \tag{31}$$

Note that (31) is spectral power ratio of the operator. In our experiment, we observed that instantaneous power of the beamformer outputs at the same frame l and frequency ω, i.e., $\widetilde{S}(l,\omega)$ and $\widetilde{C}(l,\omega)$, have shown a discernable difference with respect to the dominant voice. In Fig. 4, we illustrate it in terms of a histogram from the operator's side. This histogram was built by using 300 seconds of speech data exclusive from the testing data used for performance evaluation.

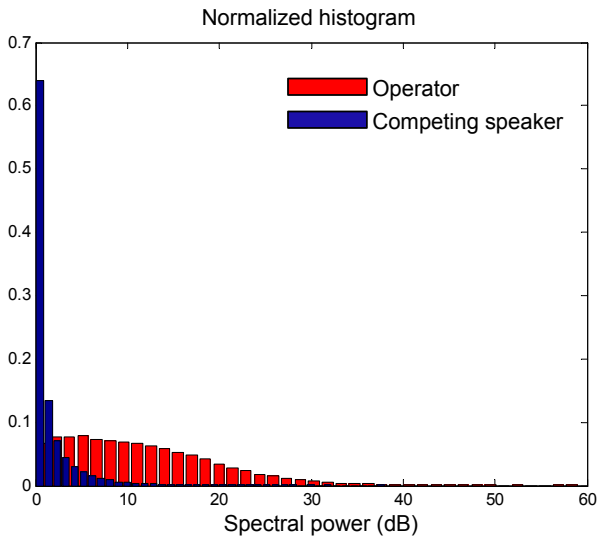

Fig. 4. Spectral power histogram

The mask for extracting operator voice from mixed source is simply presented by

$$\Gamma_S(l,w)=\begin{cases}1, & \text{if } R_S(l,\omega)>T\\ 0.01, & \text{otherwise.}\end{cases} \quad (32)$$

where T is threshold value. We decided the threshold as 2 dB based on the histogram in Fig. 4. Equation (32) means that if $R_S(l,\omega)$ is lower than T, then the corresponding time-frequency bin (l,ω) contains the competing speaker's voice signal. Thus the signal at (l,ω) should be attenuated by 0.01.

The mask for competing speaker's voice is made via similar procedure through (31) and (32), and $\Gamma_C(l,w)$ denotes the mask. Finally, the separated sources by using the proposed algorithm can be stated as

$$\begin{aligned}\begin{bmatrix}\hat{S}(l,\omega)\\ \hat{C}(l,\omega)\end{bmatrix}&=\begin{bmatrix}\Gamma_S(l,\omega) & 0\\ 0 & \Gamma_C(l,\omega)\end{bmatrix}\begin{bmatrix}\tilde{S}(l,\omega)\\ \tilde{C}(l,\omega)\end{bmatrix}\\ &=\begin{bmatrix}\Gamma_S(l,\omega) & 0\\ 0 & \Gamma_C(l,\omega)\end{bmatrix}\begin{bmatrix}\mathbf{W}_S^H(\omega)\\ \mathbf{W}_C^H(\omega)\end{bmatrix}\mathbf{Y}(l,\omega)\end{aligned} \quad (33)$$

III. Experiments

A. Settings

For the evaluation of the proposed method, a microphone array was set up indoor as shown in Figure 5 for data recording. The size of the room was $5\text{m}\times 3.8\text{m}\times 2.5\text{m}$ (length, width and height), and the microphone array was located in the middle of the room .

The 4-channel microphone array receives the input signals that consist of the stationary background noise and speech of 'person 1' and 'person 2'. The source of 'person 1' moves from -40˚ to -20˚ with respect to the center normal of the microphone array. Similarly, the source of 'person 2' moves from 30˚ to 10˚. The voice sources are located 0.3m from the horizontal center of the microphone array. Stationary office sound is used for the background noise source.

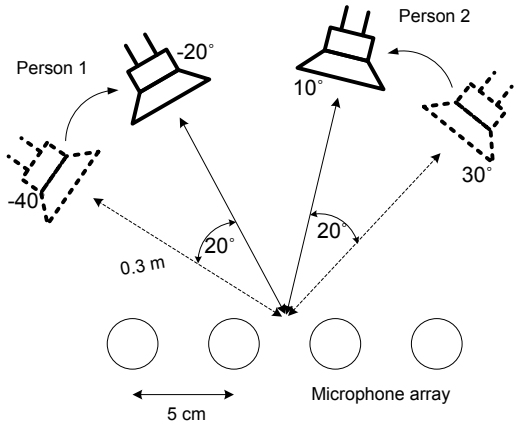

Fig. 5. Configuration of data recoding

Performance is evaluated objectively in terms of the signal enhancement factor (SEF) which is defined by

$$SEF=10\log_{10}\frac{\text{averaged power of person1's voice}}{\text{averaged power of (person2's voice + noise)}}, \quad (32)$$

The SEF measures the amount of reduction of the other person's speech signal in the processed signal. When the sound source separation is implemented successfully, the average power of the noise section decreases and the SEF increases.

B. Results

1) RTF Estimation method evaluation: Fig. 6 shows the waveforms of the input and the matched beamformer output signals with respect to the RTF estimation method. We compared three methods, namely the BLS [3], the LMS [4], and the proposed SLS method. In this experiment, both the BLS and SLS methods initialize the estimated RTF when the source moves. Comparing (c), (d), and (e) of Fig. 6, it is clear that the proposed method is more effective than the conventional methods of separating the sound sources. The BLS requires a temporal span of input signal for its initialization process in estimating the RTF, while it cannot update the changes in RTF as speaker moves. The proposed SLS method has shown that only two frames are necessary to track the change associated with speaker movements. The SLS implementation performed better in residual noise reduction compared to the other conventional methods. The LMS method performed better than the BLS in noise reduction, and exhibited similar performance in terms of waveform preservation. However, in Table 1 the SLS method achieves the best performance in terms of the SEF.

Table 1. SEF comparison

Method	SEF (dB)
BLS (baseline)	10.59
LMS	17.63
SLS (proposed)	18.98

2) TFM simulation: Fig. 7 shows the waveforms of the input, the matched beamformer output, and 'matched beamformer' + TFM. We use real recording data for evaluation of the source separation by using TFM. In Fig. 7

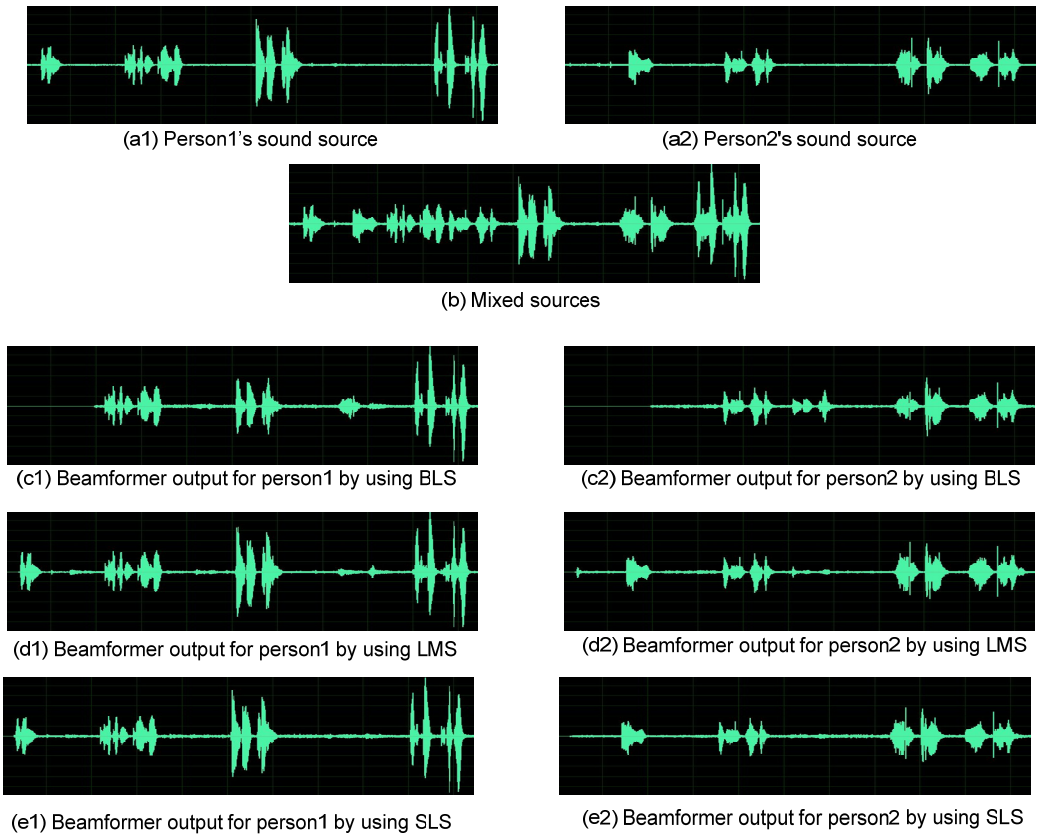

Fig. 6. Waveform comparisons

comparing (b) and (c), we can see the TFM reduces further the residual noise caused by 'person2'.

Another real data that the proposed algorithm is applied can be downloaded from http://www.umiacs.umd.edu/~jhbeh.

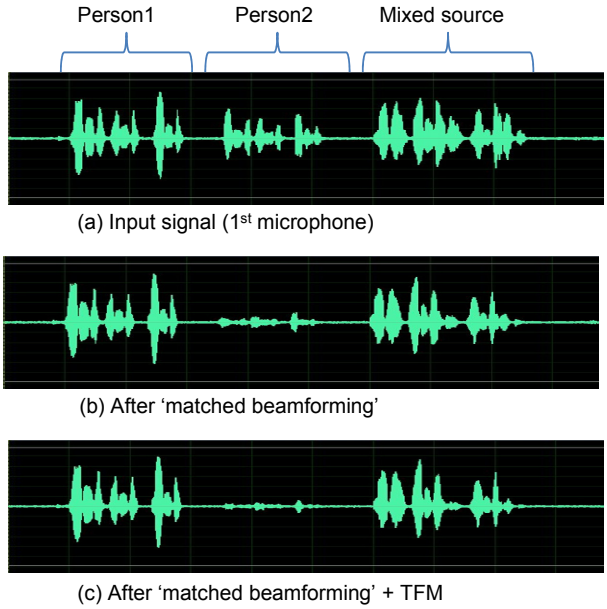

Fig. 7. Waveform of real data for evaluation and results

IV. Conclusions

We have presented the beamforming based sound source separation algorithm which operates under an assumption that the angular locations of speakers are known. In addition, we proposed the SLS method to estimate RTF for the matched beamforming. The SLS method recursively updates RTF estimate with current input signal, thus it requires minimal initialization period in data collection and results in significant reduction in memory requirement. In addition to these advantages over the conventional estimation methods, it is shown that the proposed SLS improves performance of the matched beamforming in terms of SEF. Waveform has shown that spectral ratio based TFM can further reduce the residual voice of other party in the beamformer output voice of desired speaker.

References

[1] O. Yilmaz, S. Rickard, "Blind separation of speech mixture via time-frequency masking," *IEEE Trans. Signal Processing*., vol.52, no. 7, July, 2004.

[2] J. Beh, T. Lee, S. Ahn, H. Kim, D. Han, and H. Ko, "Enabling Directional Human-Robot Speech Interface via Adaptive Beamforming and Spatial Noise Reduction", *Proc. of IROS 2007*, pp. 3454-3459, Oct. 28 – Nov. 2, 2007.

[3] S. Gannot, D. Burshtein, and E. Weinstein, "Signal Enhancement Using Beamforming and Nonstationarity with applications to Speech," *IEEE Trans. Signal Processing*, vol. 49, no. 8, Aug, 2001.

[4] I. Cohen, "Relative Transfer Function Identification Using Speech Signals," *IEEE Trans. Speech and Audio Processing*, vol.12, no. 5, September, 2004

[5] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant room," in *Microphone Arrays*, M. Brandstein and D.Ward, Eds, Springer, 2001.

[6] L. L. Sharf, *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis*, pp.384-386, Addison-Wesley, 1991.

[7] G. Reuven, S. Gannot and I. Cohen, "Dual Source Transfer-Function Generalized Sidelobe Canceller", IEEE Trans. Audio, Speech and Language Processing, vol.16, Issue 4, pp. 717-727,May, 2008.