Journal of Science and Technology Volume 48, Issue 4, 2010 pp. 109-119

BLIND MULTI-CHANNEL SPEECH SEPARATION USING SPATIAL ESTIMATION IN TWO-SPEAKER ENVIRONMENTS

HAI QUANG DAM

ABSTRACT

This paper investigates the problem of speech separation from a mixture of two speech signals without source localization information in a room environment. Due to the lack of source information, the use of spatial detector comes at an expense of permutation ambiguity. To solve the problem, a permutation alignment algorithm based on correlation is employed to group the beamformer outputs into the correct sources. 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.

1. INTRODUCTION

In recent year, microphone arrays have seen increasing application for the acquisition of speech in hand-free, distant-talker scenarios. Based on beamforming, microphone arrays are especially promising system in term of interference reduction. These systems can be used to reduce noise in hearing aids, teleconferencing systems, hands free microphones in automobiles, computer terminals, speaker phones, and speech recognition systems. Multichannel optimum filtering requires statistical knowledge about the noise statistics, the environment and the source statistics. The beamformer coefficients are optimized in such a manner that a focused beam is steered to a desired source direction, whilst suppressing the contributions coming from other directions [1, 2]. The filter weights are designed using the information about the location of the target signal and the array geometry. From those parameters, a spatial, spectral and temporal filter are formed to match the beamforming requirement [3, 4].

Most of the beamformers considered so far require information about the desired source spatial correlation matrix. This information, however, may not be readily available especially when the source is spatially non-stationary. Hence, the estimation is performed blindly without possessing information each source, such as its location and active time. This paper considers the case of separating a speech mixture where the desired source spatial correlation matrix is not available, see Fig. 1. Thus, a spatial detector is proposed for estimating this information based on the noisy received data. Briefly, the proposed spatial detector employs the principle component analysis (PCA) to obtain the desired source subspace and consequently estimates the sources spatial information. An optimum beamformer is then developed based on these spatial information.

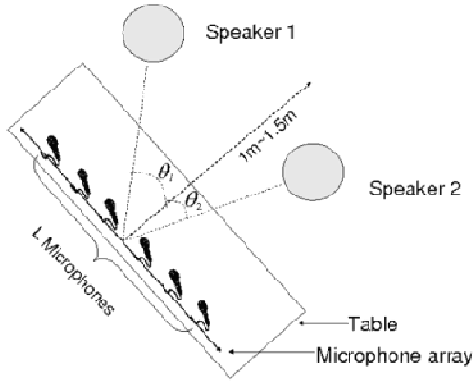

Figure 1. Position of two speakers and the microphone array in the two dimensional space

The performance of the proposed beamformer will be compared with the conventional second-order blind signal separation approach and the MVDR beamformer with calibration.

2. PROBLEM FORMULATION

Denote by $x(n)$ a $L \times 1$ discrete-time vector of the observed signal, see Fig. 1. For simplicity, we concentrate on an "offline" situation with only two speech signals. The approach in this paper, however, can be extended to a general case with more than two sources.

The observed signal $x(n)$ can be expressed as

$$\mathbf{x}(n) = \mathbf{s}_1(n) + \mathbf{s}_2(n) \tag{1}$$

where $s_1(n)$ and $s_2(n)$ are $L \times 1$ discrete-time vectors from the first and the second speech sources, respectively, at the time index n. In the frequency domain, the observed signal can be written as

$$\mathbf{x}(\omega, k) = \mathbf{s}_1(\omega, k) + \mathbf{s}_2(\omega, k) \tag{2}$$

where $x(\omega, k)$, $s_1(\omega, k)$ and $s_2(\omega, k)$ are the contribution from the observed signal, the first and the second speech sources, respectively. The objective is to separate the speech signals from the observed signal. As such, one speech source is treated as a desired source while the other becomes an undesired source. In this case, the VAD cannot be employed to detect the desired source active or inactive periods because both sources are speech signals. In the following, a source spatial detector is proposed to estimate the source spatial information based on the statistics of the observed signal.

3. SPATIAL DETECTION OF SPEECH SOURCES

Let us divide the sequence of observed signal into Q blocks, each consisting of N samples with the index $[(q-1)N+1, qN], 1 \leq q \leq Q$. The estimated correlation matrix $R_x(\omega, q)$ of the observed signal in the q^{th} block can be obtained as

$$\mathbf{R}_x(\omega, q) = \frac{1}{N} \sum_{k=(q-1)N+1}^{qN} \mathbf{x}(\omega, k)\mathbf{x}^H(\omega, k). \tag{3}$$

By assuming that the speech signals are statistically independent, the matrix $\boldsymbol{R_x(\omega, q)}$ can be decomposed as

$$\mathbf{R}_x(\omega, q) = \mathbf{R}_1(\omega, q) + \mathbf{R}_2(\omega, q) \tag{4}$$

where $\boldsymbol{R_1(\omega, q)}$ and $\boldsymbol{R_2(\omega, q)}$ are the correlation matrices for the first and the second speech signals, respectively. We have

$$\boldsymbol{R_x(\omega, q) = p_1(\omega, p)\overline{R_1}(\omega)} \tag{5}$$

where $\boldsymbol{p_1(\omega, p)}$, $\boldsymbol{p_2(\omega, p)}$ and $\boldsymbol{\overline{R_1}(\omega)}$, $\boldsymbol{\overline{R_2}(\omega)}$ are, respectively, the PSD and the spatial correlation matrices of the first and the second speech signals.

Denote by $\boldsymbol{R(\omega)}$ the estimated correlation matrix of the observed signal for $\boldsymbol{Q}$ blocks. This matrix can be obtained based on $\boldsymbol{R_x(w, q)}$ as

$$\boldsymbol{R(\omega) = \frac{1}{QN} \sum_{k=1}^{QN} x(\omega, k)x^H(\omega, k)} = \frac{1}{Q} \tag{6}$$

Clearly, during the conversation either speech sources can be active and nonactive. Therefore, there exists periods in which both speech sources are nonactive. Since $\boldsymbol{R(\omega)}$ in (6) is the average of all the estimated correlation matrices $\boldsymbol{R_x(\omega, q)}$, this matrix can be used to detect non-speech blocks or blocks with low speech presence. Thus, we propose to use a threshold $\boldsymbol{\varepsilon R(\omega, \ell, \ell)}$ to detect the speech presence where $\boldsymbol{\varepsilon}$ is a preset tolerance, $\boldsymbol{0 < \varepsilon < 1}$, and $\boldsymbol{\ell}$ is a reference microphone. The value $\boldsymbol{R(\omega, \ell, \ell)}$ is the $\boldsymbol{(\ell, \ell)^{th}}$ element of the matrix $\boldsymbol{R(\omega)}$.

Denote by $\boldsymbol{S}$ the index of all the blocks with at least one active speech source. Based on the proposed threshold, this set can be obtained as

$$\boldsymbol{S = \{q, 1 \leq q \leq Q: R_x(\omega, q, \ell,} \tag{7}$$

where $\boldsymbol{R_x(\omega, q, \ell, \ell)}$ is the $\boldsymbol{(\ell, \ell)^{th}}$ element of the matrix $\boldsymbol{R_x(\omega, q)}$. Note that, $\boldsymbol{S}$ is not an empty set since $\boldsymbol{R(\omega, \ell, \ell)}$ is the average of $\boldsymbol{R_x(\omega, q, \ell, \ell)}$, see (6).

For each $\boldsymbol{q \in S}$, denote by $\boldsymbol{R_x(\omega, p)}$ the normalized correlation matrix of the $\boldsymbol{q^{th}}$ block

. (8)

Since the $\boldsymbol{(\ell, \ell)^{th}}$ elements of the normalized spatial correlation matrices $\boldsymbol{\overline{R_1}(\omega)}$ and $\boldsymbol{\overline{R_2}(\omega)}$ are one, it follows from (5) that (8) can be rewritten as

$$\boldsymbol{\overline{R_x}(\omega, q) = \frac{p_1(\omega, q)}{p_1(\omega, q) + p_2(\omega, q)} \overline{R_1}(\omega) + \frac{p_2(\omega,}{p_1(\omega, q) +} \tag{9}$$

This equation can then be expressed as

$$\boldsymbol{\overline{R_x}(\omega, p) = \gamma_1(\omega, p)\, \overline{R_1}(\omega)} \tag{10}$$

where the values $\boldsymbol{\gamma_1(\omega, p)}$ and $\boldsymbol{\gamma_2(\omega, p)}$ represent, respectively, the proportions of the matrices $\boldsymbol{\overline{R_1}(\omega)}$ and $\boldsymbol{\overline{R_2}(\omega)}$ in the normalized correlation matrix $\boldsymbol{\overline{R_x}(\omega, p)}$, i.e.,

(11)

and

(12)

Since $p_1(\omega, q) \geq 0$ and $p_2(\omega, q) \geq 0$, we have

$$\gamma_1(\omega, q), \gamma_2(\omega, q) \geq 0 \tag{13}$$

and

. (14)

In the sequel, a spatial detection method using the subspace approach is proposed.

3.1. Spatial Detection using Signal Subspace Approach

Signal subspace is employed in a variety of signal processing applications including spectral estimation and direction of arrival (DOA) estimation [5]. In signal subspace processing, the noisy speech signal is projected into the "desired signal" or the "noise" subspace. Parameter estimation can be made by retaining only the components in the desired signal or the noise subspace.

Based on the idea of the subspace approach, the observed signal $x(\omega, k)$ is projected into either the desired or undesired signal subspace by using a $L \times 1$ vector $\mathbf{g}(\omega)$. To avoid the scaling problem, the vector vector $\mathbf{g}(\omega)$ is constrained to

$$\|\mathbf{g}(\omega)\|_F = 1. \tag{15}$$

Denote by $G_1(\omega)$ and $G_2(\omega)$ the power corresponding with the spatial correlation matrices $\overline{\boldsymbol{R}_1}(\omega)$ and $\overline{\boldsymbol{R}_2}(\omega)$, respectively,

$$G_1(\omega) = \mathbf{g}^H(\omega)\overline{\boldsymbol{R}_1}(\omega)\mathbf{g}(\omega) \tag{16}$$

and

. (17)

These values can be viewed as the projection gains for the first and second sources. Since $\overline{\boldsymbol{R}_1}(\omega)$ and $\overline{\boldsymbol{R}_2}(\omega)$ are spatial correlation matrices, these values are nonnegative.

We now investigate the condition on $G_1(\omega)$ and $G_2(\omega)$ so that the projection focuses on one of the sources, i.e., either the desired or undesired signal subspace. If

$$G_1(\omega) = G_2(\omega) \tag{18}$$

Then

$$\frac{G_1(\omega)}{G_2(\omega)} = \frac{G_2(\omega)}{G_1(\omega)} = 1 \text{ or } G_1(\omega) = G_2(\omega) = 0 \tag{19}$$

Thus, the projection does not focus on any signal subspace. On the other hand, if

(20)

then we have

$$G_1(\omega) \gg G \tag{21}$$

Without loss of generality, assume that

$$G_1(\omega) > G_2(\omega) \tag{22}$$

The case $G_1(\omega) < G_2(\omega)$ can be dealed with similarly. From (10) we have

$$\mathbf{g}^H(\omega)\overline{\boldsymbol{R}_x}(\omega,q)\mathbf{g}(\omega) = \gamma_1(\omega,q)\mathbf{g}^H(\omega)\overline{\boldsymbol{R}_1}(\omega)\mathbf{g}(\omega) + \gamma_2(\omega,q)\,\mathbf{g}^H(\omega)\overline{\boldsymbol{R}_2}(\omega)\mathbf{g}(\omega) \tag{23}$$

Or

$$\begin{aligned}\mathbf{g}^H(\omega)\overline{\boldsymbol{R}_x}(\omega,q)\mathbf{g}(\omega) &= \gamma_1(\omega,q)G_1(\omega) + \gamma_2(\omega,q)G_2(\omega)\\ &= \gamma_1(\omega,q)[G_1(\omega) - G_2(\omega)] + G_2(\omega)\end{aligned} \tag{24}$$

Note that the term $\mathbf{g}^H(\omega)\overline{\boldsymbol{R}_x}(\omega,q)\mathbf{g}(\omega)$ can also be viewed as the output-to-input power ratio of the projection for the observed signal during the q^{th} block as the input signal power is $p_1(\omega,q) + p_2(\omega,q)$ and the output power is $\mathbf{g}^H(\omega)\overline{\boldsymbol{R}_x}(\omega,q)\mathbf{g}(\omega)$. Following from (14), (22) and (24), a large value of $\gamma_1(\omega,p)$ results in a large value of $\mathbf{g}^H(\omega)\overline{\boldsymbol{R}_x}(\omega,q)\mathbf{g}(\omega)$ while a large value of $\gamma_2(\omega,p)$ gives a low value of $\mathbf{g}^H(\omega)\overline{\boldsymbol{R}_x}(\omega,q)\mathbf{g}(\omega)$. As such, the values of $\mathbf{g}^H(\omega)\overline{\boldsymbol{R}_x}(\omega,q)\mathbf{g}(\omega)$ for all $q \in S$ can be used for detecting the blocks with high proportion of each source. In other words, blocks with high proportion of the first source have high values of $\mathbf{g}^H(\omega)\overline{\boldsymbol{R}_x}(\omega,q)\mathbf{g}(\omega)$. while blocks with high proportion of the second source have low values of $\mathbf{g}^H(\omega)\overline{\boldsymbol{R}_x}(\omega,q)\mathbf{g}(\omega)$.

The task now is to find a projection vector $\mathbf{g}(\omega)$ such that the projecting subspace focuses into one of the sources. Next, the principal component analysis (PCA) technique is employed to obtain $\mathbf{g}(\omega)$.

3.2. Principle Component Analysis for Projection Vector

Principal component analysis (PCA) is a well-known technique of multi-variance analysis. This technique gives principal components that are by definition uncorrelated from a data set [6]. In blind source separation applications, PCA technique is implemented separately or as a pre-processor in combination with other techniques such as independent component analysis (ICA).

The PCA is obtained based on the covariance matrix of the data set $\boldsymbol{R}(\omega)$. From (3), the correlation matrix $\boldsymbol{R}(\omega)$ is Hermitian and can be decomposed as

(25)

where $\boldsymbol{V}(\omega)$ is an orthonormal matrix consisting of eigenvectors of $\boldsymbol{R}(\omega)$ and $\boldsymbol{\lambda}(\omega)$ is a diagonal matrix containing the corresponding eigenvalues. Thus, the projection vector $\mathbf{g}(\omega)$ is chosen as

(26)

where $v_{max}(\omega)$is the eigenvector corresponding with the maximum eigenvalue of the correlation matrix $\boldsymbol{R}(\omega)$. In the following, $\mathbf{g}(\omega)$ will be used to determine the blocks with high proportion of either the first or the second sources.

3.3. Spatial Source Detection

As discussed earlier, blocks with highest values of $\mathbf{g}^H(\omega)\overline{\boldsymbol{R}_x}(\omega,q)\mathbf{g}(\omega)$ have high contribution from the first source while the blocks with lowest values of $\mathbf{g}^H(\omega)\overline{\boldsymbol{R}_x}(\omega,q)\mathbf{g}(\omega)$ have high contribution from the second source. Thus, we propose to estimate the spatial correlation matrices for the first and the second sources by taking the average of the estimated normalized correlation matrices corresponding to $\boldsymbol{l}$ blocks with highest or lowest values of

$\mathbf{g}^H(\omega)\overline{\boldsymbol{R}_x}(\omega,q)\mathbf{g}(\omega)$. The value I is chosen smaller than half of the number of elements in $\mathcal{S}$. The average is employed to reduce the estimation error which can occur due to limited number of samples in each block.

Denote by $q_{1,1},\ldots,q_{1,I}$ and $q_{2,1},\ldots,q_{2,I}$ the index of I blocks corresponding, respectively, with I highest and lowest values of $\mathbf{g}^H(\omega)\overline{\boldsymbol{R}_x}(\omega,q)\mathbf{g}(\omega)$. The spatial correlation matrix $\hat{\bar{R}}_1(\omega)$ for the first source can be estimated based on $\overline{\boldsymbol{R}_x}(\omega,q_{1,1}),\ldots,\overline{\boldsymbol{R}_x}(\omega,q_{1,I})$ as

$$\hat{\bar{R}}_1(\omega)=\frac{1}{I}\sum_{i=1}^{I}\overline{\boldsymbol{R}_x}(\omega,q_{1,i}) \quad . \tag{27}$$

The spatial correlation matrix $\hat{\bar{R}}_2(\omega)$ for the second source can be estimated based on $\overline{\boldsymbol{R}_x}(\omega,q_{2,1}),\ldots,\overline{\boldsymbol{R}_x}(\omega,q_{2,I})$ as

$$\hat{\bar{R}}_2(\omega)=\frac{1}{I}\sum_{i=1}^{I}\overline{\boldsymbol{R}_x}(\omega,q_{2,i}) \tag{28}$$

Note that in the case where the difference between the maximum and minimum values of

$$\mathbf{g}^H(\omega)\overline{\boldsymbol{R}_x}(\omega,q)\mathbf{g}(\omega)$$

is small, then $\mathbf{g}(\omega)$ is chosen as the eigenvector corresponding with the largest eigenvalue that results in a large difference between these two values. As long as the two speech sources are in different positions in space, the two corresponding spatial correlation matrices $\overline{\boldsymbol{R}_1}(\omega)$ and $\overline{\boldsymbol{R}_2}(\omega)$ are different, and hence this task is always feasible.

Based on the fact that there are only two speakers in this context and each of the speakers has the active and nonactive time during their conversation. As such, block length \$N\$ can be chosen low enough for obtaining the one-speech blocks from the observed signal. Thus, the proportion of the non-dominated source in the matrices $\hat{\bar{R}}_1(\omega)$ and $\hat{\bar{R}}_2(\omega)$ is approximately equal to zeros and this proportion can be neglected. These matrices are now used to estimate the optimum beamformer in each frequency bin.

4. OPTIMUM BEAMFORMER USING SPATIAL INFORMATION

Based on the estimated spatial correlation matrices $\hat{\bar{R}}_1(\omega)$ and $\hat{\bar{R}}_2(\omega)$, an optimum beamformer is obtained for each frequency bin ω. Denote by $w_1(\omega)$ the beamformer weight for the first source. This weight vector can be obtained by solving the optimization problem

$$\begin{cases}\min w_1^H(\omega)\hat{\bar{R}}_2(\omega)w_1(\omega)\\ \text{subject to } w_1^H(\omega)\hat{\bar{d}}_1(\omega)=1\end{cases} \tag{29}$$

where $\hat{\bar{d}}_1(\omega)$ is the estimated cross-correlation vector between the first source and a ℓ^{th} reference microphone. This vector is also the ℓ^{th}column of the matrix $\hat{\bar{R}}_1(\omega)$. Similarly, the beamformer weight $w_2(\omega)$ for the second source can be obtained as the solution to the optimization problem

$$\begin{cases}\min w_2^H(\omega)\hat{\bar{R}}_1(\omega)w_2(\omega)\\ \text{subject to } w_2^H(\omega)\hat{\bar{d}}_2(\omega)=1\end{cases} \tag{30}$$

where $\hat{\bar{d}}_2(\omega)$ is the i^{th} column of the matrix $\hat{\bar{R}}_2(\omega)$ The solutions to these optimization problems can be expressed as

$$w_1(\omega) = \frac{\hat{\bar{R}}_2^{-1}(\omega)\hat{\bar{d}}_1(\omega)}{\hat{\bar{d}}_1^{H}(\omega)\hat{\bar{R}}_2^{-1}(\omega)\hat{\bar{d}}_1(\omega)} \tag{31}$$

and

$$w_2(\omega) = \frac{\hat{\bar{R}}_1^{-1}(\omega)\hat{\bar{d}}_2(\omega)}{\hat{\bar{d}}_2^{H}(\omega)\hat{\bar{R}}_1^{-1}(\omega)\hat{\bar{d}}_2(\omega)} \tag{32}$$

The beamformer outputs for the two sources are calculated as

(33)

and

. (34)

The remaining problem is to align the beamformer output in different frequency bins to the same source. This problem is also referred to as permutation alignment problem as there is permutation ambiguity in the solution. In the sequel, the correlation between the beamformer outputs in neighboring frequencies is employed to overcome the permutation problem.

5. PERMUTATION ALIGNMENT

The permutation alignment problem can be overcome by using correlation, localization or source statistic property [7]. The correlation approach is preciseness. This approach, however, might not be robust as a misalignment at one frequency affects the results of other frequencies. Since we consider only two speech sources, the correlation approach is chosen for the permutation alignment. As such, permutation decision is based on inter-frequency correlation of the output signal amplitudes. This is done based on the assumption that the amplitudes of the output signals from the same source for adjoining frequencies are correlated.

The permutation alignment can be performed continuously from a reference frequency. In this case, this frequency is chosen in the middle of the frequency range. Permutation correlation is then performed in two directions, with increasing and decreasing frequency indexes until the end of the frequency range. More specifically, for two neighboring frequencies ω_i and ω_{i+1}, the following correlations between the two beamformer outputs are obtained

$$con_{1,1} = \frac{\mu(|y_1(\omega_i,k)y_1(\omega_{i+1},k)|) - \mu(|y_1(\omega_i,k)|)\mu(|y_1(\omega_{i+1},k)|)}{\sigma(|y_1(\omega_i,k)|)\sigma(|y_1(\omega_{i+1},k)|)} \tag{35}$$

$$con_{1,2} = \frac{\mu(|y_1(\omega_i,k)y_2(\omega_{i+1},k)|) - \mu(|y_1(\omega_i,k)|)\mu(|y_2(\omega_{i+1},k)|)}{\sigma(|y_1(\omega_i,k)|)\sigma(|y_2(\omega_{i+1},k)|)} \tag{36}$$

$$con_{2,1} = \frac{\mu(|y_2(\omega_i,k)y_1(\omega_{i+1},k)|) - \mu(|y_2(\omega_i,k)|)\mu(|y_1(\omega_{i+1},k)|)}{\sigma(|y_2(\omega_i,k)|)\sigma(|y_1(\omega_{i+1},k)|)} \tag{37}$$

and

$$con_{2,2} = \frac{\mu(|y_2(\omega_i,k)y_2(\omega_{i+1},k)|) - \mu(|y_2(\omega_i,k)|)\mu(|y_2(\omega_{i+1},k)|)}{\sigma(|y_2(\omega_i,k)|)\sigma(|y_2(\omega_{i+1},k)|)}. \quad (38)$$

Here $|\cdot|$ is the absolute operation while $\mu(\cdot)$ and $\sigma(\cdot)$ are, respectively, the mean and the standard deviation of $(\cdot)$. Permutation alignment is performed if the following equation is satisfied

$$con_{1,2}+con_{2,1}>con_{1,1}+con_{2,2}\ . \quad (39)$$

Finally, two frequency domain output signals are passed through the synthesis filterbank to obtain the time domain output representations.

6. EVALUATIONS

Measurements and evaluations have been performed in a real room environment using a linear microphone array consisting of 6 microphones. The distance between two adjacent microphones is 6 cm. The positions of the two speakers are shown in Fig. 1 with $\theta_1 = 50^\circ$ and $\theta_2 = 20^\circ$. The value N was chosen as the number of samples in 1 s period while I and ε were chosen as 5 and 0.1, respectively.

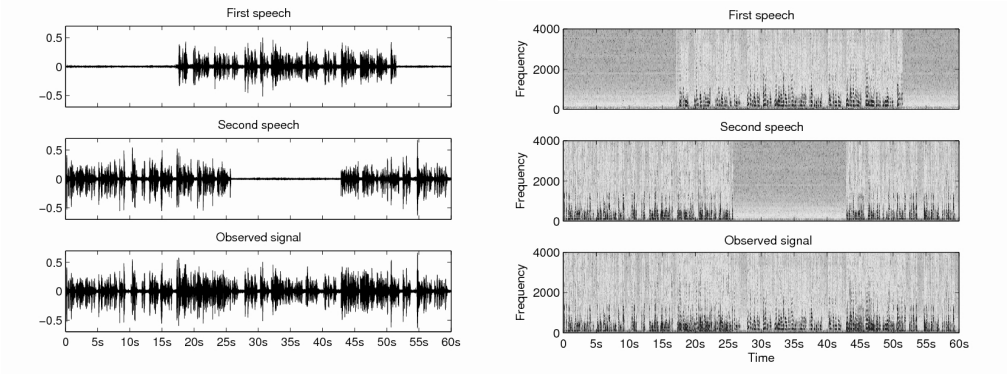

Figure 2. Time domain plots and spectograms of the original speech signals and the obversed signal at the 4th microphone

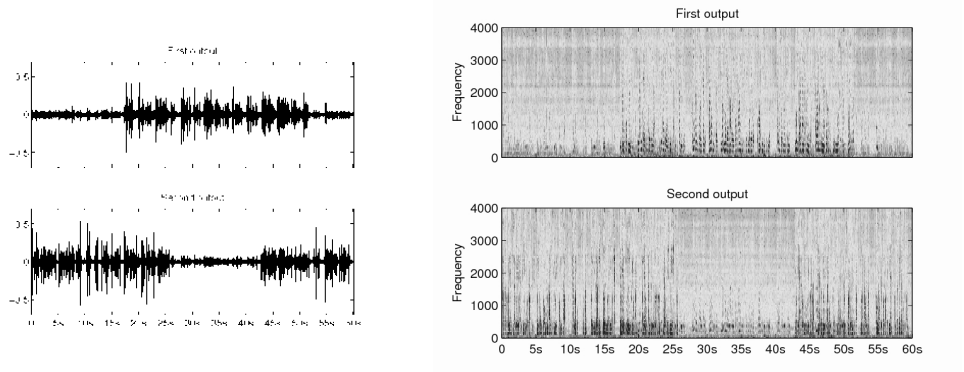

Figure 3. Time domain plots and spectograms of the proposed beamformer outputs

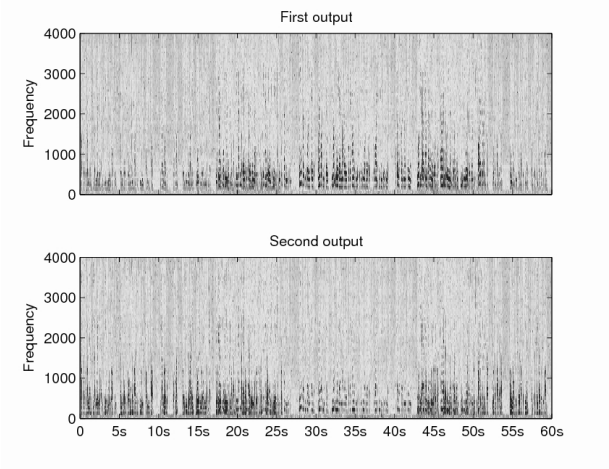

Figure 4. Spectograms of two speech dominant outputs of the second-order gradient base BSS algorithm

To quantify the performance of the proposed beamformer, the interference suppression (IS) and source distortion (SD) measures in [8] are employed. Here, one speaker is viewed as the desired signal while the other is the undesired or interference signal.

The proposed beamformer is performed in the frequency domain with the same parameters as in [8]. Its performance is compared with the gradient based second-order blind signal separation (BSS) algorithm and the MVDR beamformer using calibrated information [9]. Note that for the case with two speech sources, the MVDR beamformer with calibration is also a fixed optimum beamformer using calibration which aims to minimize the average undesired source output power.

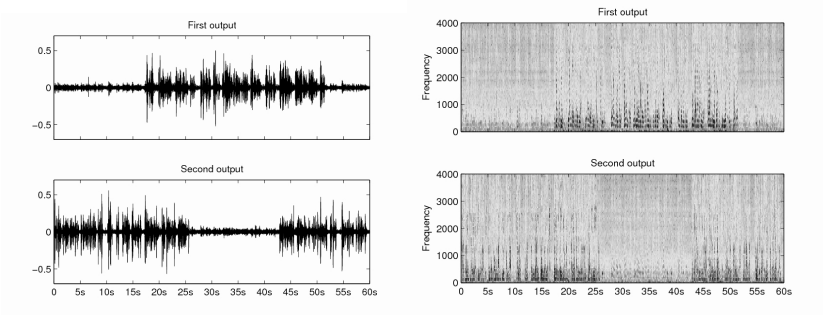

Figure 5. Time domain plots and spectograms for the outputs of the MVDR beamformmer with calibrated information

Figure 2 shows time domain plots and spectrograms of the two speech signals and the observed signal. The duration of the observed time is 60~s. The speech signals from the two speakers occur at different times and can overlap with each other in the observed signal.

Table 1. The interference suppression and the source distortion levels in the outputs of the proposed methods, the second-order gradient based BSS and the MVDR using calibrated information

Methods	First Output		Second Output	
	IS(dB)	SD(dB)	IS(dB)	SD(dB)
Proposed beamformer	11.2	-28.9	10.8	-27.5
Second-order BSS	7.3	-25.5	7.17	-26.3
MVDR with calibrated information	13.8	-30	11.9	-7.6

Figures 3 depict, time domain plots and spectrograms of the proposed beamformer outputs with permutation alignment. The first output is the speech signal from the first speaker while the second output is from the second speaker. Thus, the proposed beamformer can separate the two speech signals from the observed mixture. Informal listening tests suggest good quality speech signal outputs from the propose structure.

The time domain plots and the spectrograms of the second order BSS and the MVDR beamformer with calibrated information are depicted in Figs. 4, 5. For the BSS algorithm, the number of output is the same with the number of microphones. Both beamformers recover the two speech signals with low distortion.

Table 1 shows the IS and SD levels for the two outputs of the proposed beamformer, the second-order gradient based BSS and the MVDR beamformer using calibration. The table shows an improvement in the IS and SD levels of the proposed beamformer when compared with the gradient based second-order BSS. More specifically, the proposed method improves approximately $3-4dB$ in the IS measures and $1-3dB$ in the SD measures over the second-order BSS.

The IS levels of the proposed beamformer is slightly lower than that of the MVDR beamformer using calibration. On the other hand, the SD measures of the proposed beamformer the MVDR beamformer remain approximately the same. Thus, the source spatial correlation matrices estimated by using the proposed spatial detector closely match with those obtained by using calibration.

7. CONCLUSION

In this paper, a spatial detector in multi-speaker environment is developed for the case where source localization or calibration information is not available. An optimum beamformer with permutation alignment is then proposed with includes an estimation of the source spatial correlation matrices. The performance of the proposed beamformer is then compared with the second-order gradient based BSS and the MVDR beamformer with calibration information using real data. Simulation results show an improvement of the IS and SD levels of the proposed beamformer over the second-order BSS algorithm. In addition, the IS levels of the proposed beamformer are slightly lower than that of the MVDR beamformer while the SD levels remain

approximately the same. Thus, the spatial correlation matrices estimated by using the proposed detector closely match with that of the MVDR beamformer using calibration.

REFERENCES

1. M. Brandstein and D. Ward (Eds.) - Microphone Arrays: Signal Processing Techniques and Applications, Springer-Verlag, 2001.
2. S. Nordebo, I. Claesson, and S. Nordholm - Adaptive beamforming: Spatial filter designed blocking matrix, IEEE Journal of Oceanic Engineering **19** (1994) 583-590.
3. H. Q. Dam, S Nordholm, H. H Dam, and S. Y. Low - Postfiltering using multichannel spectral estimation in multispeaker enviroments, EURASIP Journal on Advances in Signal Processing **ID 860360** (2008) 1-10.
4. N. Grbic, S. Nordholm, and A. Cantoni - Optimal fir subband beamforming for speech enhancement in multipath enviroments, IEEE Signal Processing Letters **10** (11) (2003) 335-338.
5. F. Jabloun and B. Champagne - Speech Enhancement, chapter Signal Subspace Techniques for Speech Enhancement, pp. 135-139, Springer-Verlag, 2005.
6. T. Jolliffe - Principal Component Analysis, Springer-Verlag, 2 edition, 2002.
7. H. Sawada, R. Mukai, S. Araki, and S. Makino - Speech Enhancement, chapter Frequency-Domain Blind Source Separation, pp. 299-328, Springer-Verlag, 2005.
8. H. Q. Dam, S. Nordholm, H. H Dam, and S. Y. Low - Adaptive beamformer for hand-free communication system in noisy enviroments, IEEE Int. Symposium on Circuits and Systems **2** (2005) 856-859.
9. J. Benesty, S. Makino, and J. Chen (Eds.) - Speech Enhancement, Springer-Verlag, 2005.

Address:

Faculty of Information Technology,

University of Science, VNU – HCMC.

Received June 16, 2010