World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:10, No:8, 2016

Blind Speech Separation Using SRP-PHAT Localization and Optimal Beamformer in Two-Speaker Environments

Hai Quang Hong Dam, Hai Ho, Minh Hoang Le Ngo

International Science Index, Computer and Information Engineering Vol:10, No:8, 2016 waset.org/Publication/10005265

Abstract**—**This paper investigates the problem of blind speech separation from the speech mixture of two speakers. A voice activity detector employing the Steered Response Power - Phase Transform (SRP-PHAT) is presented for detecting the activity information of speech sources and then the desired speech signals are extracted from the speech mixture by using an optimal beamformer. For evaluation, the algorithm effectiveness, a simulation using real speech recordings had been performed in a double-talk situation where two speakers are active all the time. Evaluations show that the proposed blind speech separation algorithm offers a good interference suppression level whilst maintaining a low distortion level of the desired signal.

Keywords**—**Blind speech separation, voice activity detector, SRP-PHAT, optimal beamformer.

I. Introduction

IN recent years, research in speech separation for cocktail-party or multiple-speaker environment has been very actively conducted by using multi-channel systems like microphone arrays. Hence, this problem is very attractive in the speech enhancement area when the observed signal is obtained from several speakers in different locations. Many applications have benefited from the multi-channel speech separation techniques such as in hearing aids, multi-talker speech separation, hands- free microphones, robot systems, speaker phones, and speech recognition systems. Here, the localization information of speech sources is very important for speech separation due to the fact that all speech signals have the same spectral characteristic. In the case of localization information available, many methods have been proposed for evaluation of sources' spatial information such as evaluation using a pre-known geometry localization like the array geometry and source localization, a calibration method by using training samples of prerecording desired and undesired sources [1], [2]. Using available sources' spatial information, many separation techniques have been proposed like steering beamforming, optimal beamforming, post-filtering [3], [4]. There, optimal beamforming techniques are very popular because optimal beamformers are used to exploit the spatial information of desired and undesired signals in such a way that the desired one is extracted and undesired signals are suppressed. Then, the optimal beamformers are designed by using the spatial information to suppress the contribution of all undesired signals while preserving the contribution of the desired signal. Specifically, the optimal beamformer weights are calculated by using the knowledge about the location of the target signal and the array geometry [2], [3]. However, when the localization knowledge is not known a priori then the observed mixture signals are the only available data for the speech enhancement. In this case, blind source separation (BSS) techniques are developed for separating the different sound sources. Many blind speech separation techniques using microphone array have been proposed for the speech separation in both time domain and frequency domain. Some very popular BBS techniques for the speech separation are Independent Component Analysis (ICA), maximum likelihood, and kurtosis maximization [5]-[7]. Most of the BBS techniques are based on the independent characteristics of speech sources in the observed signal. In the case of blind separation in a multiple-speaker environment, different BSS techniques are proposed in both time domain and time-frequency domain [8]-[12].

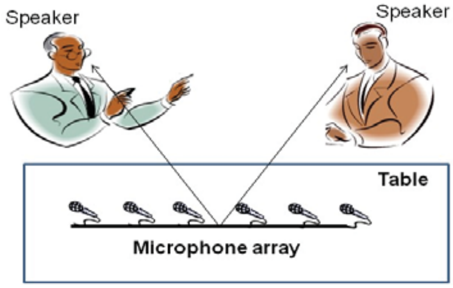

Fig. 1 Position of speakers and the microphone array in the Two-Speaker Environment

This paper considers the case of blindly separating of speech signals from a two speech sources' mixture (see Fig. 1). Thus, the source activity information is blindly estimated in the frequency domain without having prior knowledge about the location of speech sources. In [9], [10], active and inactive periods of speech sources have been detected for their spatial information. However, in some special cases, sources' inactive periods are not available like a double-talk situation where two sources are active all the time. To overcome with double- talk situation, in this paper, a voice activity detector

Dr. Hai Quang Hong Dam is a lecturer with the University of Information Technology, Ho Chi Minh City, Vietnam (Phone: 848-3957-4779; e-mail: damhai@uit.edu.vn).

M.Sc Hai Ho is a lecturer with the University of Information Technology, Ho Chi Minh City, Vietnam (e-mail: haih@uit.edu.vn).

Mr. Minh Hoang Le Ngo is a Master student with the University of Information Technology, Ho Chi Minh City, Vietnam (e-mail: ngohoangleminh@gmail.com).

International Science Index, Computer and Information Engineering Vol:10, No:8, 2016 waset.org/Publication/10005265

for both speech sources employing SRP-PHAT is presented for obtaining the speech sources' activity information. The SRP-PHAT is based on the cross-correlation and phase-transform weighting of the observed signals from all microphone pairs in the array [13]. Based on the sources' activity information, an optimal beamformer is proposed for extraction of each speech source from the observed signals. The effectiveness of the proposed algorithm had been evaluated by using the real speech recordings in double-talk situation. Evaluations in a real double-talk environment reveal a good suppression level of the undesired speech source whilst maintaining a low distortion level of the desired speech source.

II. Problem Formulation

Consider a linear microphone array with L microphones, and observed signal, in this case, is the mixing speech of two speakers sitting in front of the array (see Fig. 1). A $L\times 1$ discrete-time vector of the observed signal is denoted by $\mathbf{x}(\mathrm{n})$ (see Fig. 1). The observed signal $\mathbf{x}(\mathrm{n})$ can be expressed as

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

where $\mathbf{s}_1(\mathrm{n})$ and $\mathbf{s}_2(\mathrm{n})$ are L × 1 discrete-time vectors from the first and second sources, respectively, at the time index n. Initially, the observed signal is decomposed into the frequency domain by using a frequency analysis method. 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 $\mathbf{x}(\omega,k)$, $\mathbf{s}_1(\omega,k)$, and $\mathbf{s}_2(\omega,k)$ are the contribution from the observed signal, the first and the second speech sources, respectively in frequency bin ω. The objective is to separate the speech signals from the observed signal. As such, one speech source is treated as the desired source while the other becomes an undesired source. In this case, a voice activity detector employing SRP-PHAT is proposed to detect the speech sources' activity information based on the statistics of the observed signals.

III. Voice Activity Detector Employing SRP-PHAT

Let us divide the sequence of observed signal into Q blocks, each consisting of N samples with the index [(q−1)N+1,qN], $1 \le q \le Q$. The SRP-PHAT of the observed signal in the q^th block can be obtained as

$$\Psi_{\mathbf{x}}(\omega,q)=\frac{\sum_{n=1}^{L}\sum_{m=n+1}^{L}R_{\mathbf{x}}(\omega,q,n,m)}{p_{\mathbf{x}}(\omega,q)} \tag{3}$$

where $R_{\mathbf{x}}(\omega,q,n,m)$ is the cross-correlation between m^th and n^th microphone observed signals from the q^th block in frequency bin ω, $p_{\mathbf{x}}(\omega,q)$ is the power spectral density (PSD) of the observed signals from the q^th block in frequency bin ω. Here, the cross-correlation $R_{\mathbf{x}}(\omega,q,n,m)$ can be estimated as follows

$$R_{\mathbf{x}}(\omega,q,n,m)=\frac{1}{N}\sum_{k=(q-1)N+1}^{qN}x(\omega,k,n)x(\omega,k,m)^* \tag{4}$$

where $(.)^*$ denotes the complex conjugate operator, $x(\omega,k,n)$ and $x(\omega,k,m)$ are m^th and n^th elements of the observed vector $\mathbf{x}(\omega,k)$. The PSD of the observed signals $p_{\mathbf{x}}(\omega,q)$ can be estimated by using the observed signals at a reference microphone ℓ $(1\le \ell \le L)$ as

$$p_{\mathbf{x}}(\omega,q)=\frac{1}{N}\sum_{k=(q-1)N+1}^{qN}x(\omega,k,\ell)x(\omega,k,\ell)^* \tag{5}$$

where $x(\omega,k,\ell)$ is the ℓ^{th} element of the observed vector $\mathbf{x}(\omega,k)$.

To avoid the division by 0 in (3) i.e. periods in which all speech sources are inactive, we propose to use a threshold $\varepsilon p_{\mathrm{x}}(\omega)$ to detect the speech presence where ε is a preset tolerance, $0<\varepsilon<1$, and $p_{\mathrm{x}}(\omega)$ is the PSD of observed signals in frequency bin ω. Here, $p_{\mathrm{x}}(\omega)$ can be calculated as

$$p_{\mathbf{x}}(\omega)=\frac{1}{qN}\sum_{k=1}^{qN}x(\omega,k,\ell)x(\omega,k,\ell)^* \tag{6}$$

Denote by 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

$$S=\{q, \quad 1\le q\le Q \quad : \quad p_{\mathbf{x}}(\omega,q)\ge \varepsilon p_{\mathbf{x}}(\omega) \quad \} \tag{7}$$

Note that, S is not an empty set since $p_{\mathrm{x}}(\omega)$ is the average of all $p_{\mathrm{x}}(\omega,\mathrm{q})$, see (5) and (6). The SRP-PHAT of the first and second sources' signals in the q^th block can be obtained as

$$\Psi_{\mathbf{s1}}(\omega,q)=\frac{\sum_{n=1}^{L}\sum_{m=n+1}^{L}R_{\mathbf{s1}}(\omega,q,n,m)}{p_{\mathbf{s1}}(\omega,q)} \tag{8}$$

and

$$\Psi_{\mathbf{s2}}(\omega,q)=\frac{\sum_{n=1}^{L}\sum_{m=n+1}^{L}R_{\mathbf{s2}}(\omega,q,n,m)}{p_{\mathbf{s2}}(\omega,q)} \tag{9}$$

where $R_{\mathbf{s1}}(\omega,q,n,m)$ and $R_{\mathbf{s2}}(\omega,q,n,m)$ are cross correlation between m^th and n^th microphone for first and second speech sources from the q^th block in frequency bin ω, $p_{\mathbf{s1}}(\omega,q)$ and $p_{\mathbf{s2}}(\omega,q)$ are power spectral density (PSD) of the first and second speech sources from the q^th block in frequency bin ω. Here, cross-correlation $R_{\mathbf{s1}}(\omega,q,n,m)$ and $R_{\mathbf{s2}}(\omega,q,n,m)$ can be estimated as

$$R_{\mathbf{s1}}(\omega,q,n,m)=\frac{1}{N}\sum_{k=(q-1)N+1}^{qN}s_1(\omega,k,n)s_1(\omega,k,m)^* \tag{10}$$

and

$$R_{\mathbf{s2}}(\omega,q,n,m)=\frac{1}{N}\sum_{k=(q-1)N+1}^{qN}s_2(\omega,k,n)s_2(\omega,k,m)^* \tag{11}$$

International Science Index, Computer and Information Engineering Vol:10, No:8, 2016 waset.org/Publication/10005265

where $s_1(\omega,k,n)$ and $s_1(\omega,k,m)$ are m$^{\text{th}}$ and n$^{\text{th}}$ elements of the first source source vector $\mathrm{s}_1(\omega,k)$, and $s_2(\omega,k,n)$ and $s_2(\omega,k,m)$ are m$^{\text{th}}$ and n$^{\text{th}}$ elements of the first second source vector $\mathrm{s}_2(\omega,k)$. The PSD of the two speech source signals $p_{\mathbf{x}}(\omega,q)$ can be estimated as

$$p_{\mathbf{s1}}(\omega,q)=\frac{1}{N}\sum_{k=(q-1)N+1}^{qN} s_1(\omega,k,\ell)s_1(\omega,k,\ell)^* \tag{12}$$

and

$$p_{\mathbf{s2}}(\omega,q)=\frac{1}{N}\sum_{k=(q-1)N+1}^{qN} s_2(\omega,k,\ell)s_2(\omega,k,\ell)^* \tag{13}$$

Due to the fact that two speech signals $\mathbf{s}_1$(n) and $\mathbf{s}_2$(n) are statistically independent, the cross-correlation between them are zeros for all frequency bin. So, $R_{\mathbf{x}}(\omega,q,n,m)$ and $p_{\mathbf{x}}(\omega,q)$ can be expressed as follows

$$R_{\mathbf{x}}(\omega,q,n,m)=R_{\mathbf{s1}}(\omega,q,n,m)+R_{\mathbf{s2}}(\omega,q,n,m) \tag{14}$$

and

$$p_{\mathbf{x}}(\omega,q)=p_{\mathbf{s1}}(\omega,q)+p_{s2}(\omega,q) \tag{15}$$

From (3), (8), (9), (14), and (15), we have the following expression

$$\Psi_{\mathbf{x}}(\omega,q)=\frac{p_{\mathbf{s1}}(\omega,q)\Psi_{\mathbf{s1}}(\omega,q)+p_{\mathbf{s2}}(\omega,q)\Psi_{\mathbf{s2}}(\omega,q)}{p_{\mathbf{s1}}(\omega,q)+p_{\mathbf{s2}}(\omega,q)} \tag{16}$$

In this research, two speech sources are unmoved during the conversation, so we can use the SRP-PHAT to localize the source position in the room [12], [13]. As such, the SRP-PHAT of each speech source can be supposed to be unchanged for all block qϵS. Hence, $\Psi_{s1}(\omega,q)$ is supposed to be equal to $\Psi_{s1}(\omega)$, and $\Psi_{s2}(\omega,q)$ is supposed to be equal to $\Psi_{s2}(\omega)$ for all block qϵS. Then, (15) can be rewritten as follows

$$\Psi_{\mathbf{x}}(\omega,q)=\frac{p_{\mathbf{s1}}(\omega,q)}{p_{\mathbf{s1}}(\omega,q)+p_{\mathbf{s2}}(\omega,q)}\Psi_{\mathbf{s1}}(\omega)+ \\ \frac{p_{\mathbf{s2}}(\omega,q)}{p_{\mathbf{s1}}(\omega,q)+p_{\mathbf{s2}}(\omega,q)}\Psi_{\mathbf{s2}}(\omega) \tag{17}$$

Denote $\gamma_{s1}(\omega,q)$ as follows

$$\gamma_{\mathbf{s1}}(\omega,q)=\frac{p_{\mathbf{s1}}(\omega,q)}{p_{\mathbf{s1}}(\omega,q)+p_{\mathbf{s2}}(\omega,q)} \tag{18}$$

so, (17) can be rewritten as follows

$$\Psi_{\mathbf{x}}(\omega,q)=\gamma_{s1}(\omega,q)\Psi_{\mathbf{s1}}(\omega)+(1-\gamma_{s1}(\omega,q))\Psi_{\mathbf{s2}}(\omega) \tag{19}$$

Clearly, (18) shows the contribution balance between two speech sources in block q. Here, $\gamma_{s1}(\omega,q)$ is the contribution of the first speech source, and 1- $\gamma_{s1}(\omega,q)$ is the contribution of the second speech source. As such, during the conversation, contributions of speech sources can be changed from block to block. Blocks, in which the contribution of one speech source is dominant in comparison with the contribution of another speech source, are useful for the sources' activity estimation. In the complex plane, based on (19), the point of $\Psi_{x}(\omega,q)$ is located in the link between two points of $\Psi_{s1}(\omega)$ and $\Psi_{s2}(\omega)$. Also, the block's point located near vertices should represent the domination of one speech source, see Fig. 3. As such, block of first source domination can be detected as block q_1, and the block of second source domination q_2 can be detected as

$$q_1,q_2=\arg\max_{q_1,q_2\in S}\left|\Psi_{\mathbf{x}}(\omega,q_1)-\Psi_{\mathbf{x}}(\omega,q_2)\right| \tag{20}$$

here | • | is the absolute operation. To reduce the frequency mismatch, we can use SRP-PHAT of a frequency range [ω_1 ω_2] which can be estimated as

$$\Psi_{\mathbf{x}}([\omega_1\ \omega_2],q)=\sum_{\omega=\omega_1}^{\omega_2}\Psi_{\mathbf{x}}(\omega,q) \tag{21}$$

Signal blocks, where the contribution of one source is dominant compared to the contribution of another source, can be estimated as I blocks which SRP-PHAT are nearest to SRP-PHAT of q$_1$$^{\text{th}}$ block or SRP-PHAT of q$_2$$^{\text{th}}$ block. Then, $\prod_1$ and $\prod_2$ are proposed to be the sets of observed signal x(n) in I blocks which SRP-PHAT are nearest to SRP-PHAT of q$_1$$^{\text{th}}$ block and q$_2$$^{\text{th}}$ block, respectively. In practice, the value I can be chosen as smaller than 5% of the number of elements in S.

IV. Optimal Beamformer Using Sources' Activity Information

For extracting each speech source from the observed signals, the optimal beamforming technique is used based on the sources' activity information obtained by voice activity detector in such a way that the desired one is extracted and undesired one is suppressed. In each frequency bin ω, the correlation matrix $\mathbf{R}_1(\omega)$ for the first source can be estimated as

$$\mathbf{R}_1(\omega)=\frac{1}{IQ}\sum_{k\in\Pi_1}\mathbf{x}(\omega,k)\mathbf{x}^{\mathbf{H}}(\omega,k) \tag{22}$$

Due to the small value of I, the contribution of second speech sources in the correlation matrices $\mathbf{R}_1(\omega)$ is much smaller in comparison with the contribution of the first speech source. The correlation matrix R$_2$ (ω) for the second source can be estimated as follows

$$\mathbf{R}_2(\omega)=\frac{1}{IQ}\sum_{k\in\Pi_2}\mathbf{x}(\omega,k)\mathbf{x}^{\mathbf{H}}(\omega,k) \tag{23}$$

These matrices are now used to desire an optimal beamformer in each frequency bin. Based on the estimated sources' correlation matrices $\mathbf{R}_1(\omega)$ and $\mathbf{R}_2(\omega)$ in each

International Science Index, Computer and Information Engineering Vol:10, No:8, 2016 waset.org/Publication/10005265

frequency bin ω, an optimal beamformer is desired for each speech source in the frequency bin ω. The beamformer weight for the first source is denoted by $\mathbf{w}_1(\omega)$. This weight vector can be obtained by solving the following optimization problem

$$\begin{cases}\min \mathbf{w}_1^{\mathbf{H}}(\omega)\mathbf{R}_2(\omega)\mathbf{w}_1(\omega)\\ \text{subject to } \mathbf{w}_1^{\mathbf{H}}\mathbf{d}_1(\omega)=1\end{cases} \tag{24}$$

where $\mathbf{d}_1(\omega)$ is the estimated cross-correlation vector between the first source and a ℓ^{th} reference microphone. The vector $\mathbf{d}_1(\omega)$ is also the ℓ^{th} column of the matrix $\mathbf{R}_1(\omega)$. Similarly, the beamformer weight $\mathbf{w}_2(\omega)$ for the second source can be obtained as the solution to the optimization problem

$$\begin{cases}\min \mathbf{w}_2^{\mathbf{H}}(\omega)\mathbf{R}_1(\omega)\mathbf{w}_2(\omega)\\ \text{subject to } \mathbf{w}_2^{\mathbf{H}}\mathbf{d}_2(\omega)=1\end{cases} \tag{25}$$

where $\mathbf{d}_2(\omega)$ is the ℓ^{th} column of the matrix $\mathbf{R}_2(\omega)$. The solutions to two optimization problems can be expressed as

$$\mathbf{w}_1(\omega)=\frac{[\mathbf{R}_2(\omega)]^{-1}\mathbf{d}_1(\omega)}{\mathbf{d}_1^{\mathbf{H}}(\omega)[\mathbf{R}_2(\omega)]^{-1}\mathbf{d}_1(\omega)} \tag{26}$$

and

$$\mathbf{w}_2(\omega)=\frac{[\mathbf{R}_1(\omega)]^{-1}\mathbf{d}_2(\omega)}{\mathbf{d}_2^{\mathbf{H}}(\omega)[\mathbf{R}_1(\omega)]^{-1}\mathbf{d}_2(\omega)} \tag{27}$$

The beamformer outputs for the two sources are calculated as

$$y_1(\omega,k)=w_1^H\mathbf{x}(\omega,k) \tag{28}$$

and

$$y_2(\omega,k)=w_2^H\mathbf{x}(\omega,k) \tag{29}$$

Finally, time-domain outputs $y_1(n)$ and $y_2(n)$ of speeches from two speakers can be obtained from the frequency-domain beamformer outputs $y_1(n,k)$ and $y_2(n,k)$ by using the synthesis reconstruction.

V. Experimental Results

For performance evaluations of the proposed blind speech separation algorithm, a simulation in double-talk situation is performed in a real room environment by using a linear microphone array consisting of six microphones. Here, the distance between two adjacent microphones is 6 cm, and the positions of two speakers are shown in Fig. 1. The distances between the array and speakers are about 1m ~ 1.5m. The duration of the observed signal is 60 seconds, and the value N was chosen as the number of samples in the 0.25 s period, while ℓ, I, and ε were chosen as 4, 15 and 0.1, respectively. Fig. 2 shows time domain plots of two speech signals and the observed signal. The speech signals from two speakers occur at same times and overlap with each other in the observed signal.

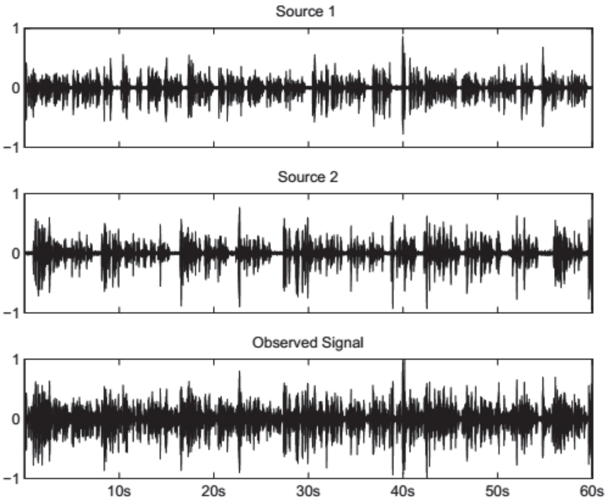

Fig. 2 Time domain plots of the original speech signals and the observed signal at the 4th microphone

TABLE I
THE INTERFERENCE SUPPRESSION AND THE SOURCE DISTORTION LEVELS IN THE OUTPUTS OF THE PROPOSED METHOD AND THE SECOND-ORDER BSS ALGORITHM

Method	First output		Second output	
	IS (dB)	SD (dB)	IS (dB)	SD (dB)
Proposed beamformer	6,9	-21,2	7,6	-23,4
Second-order BSS	2,1	-20,5	1,9	-22,3

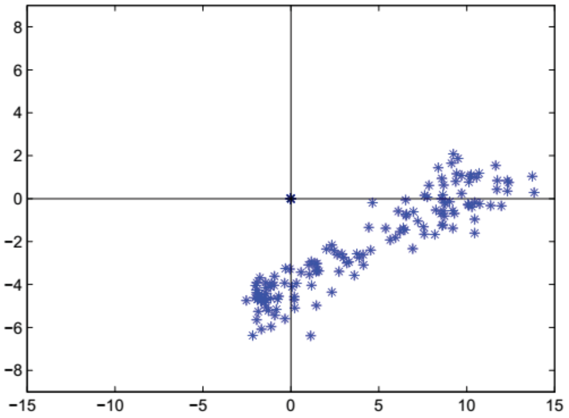

Fig. 3 Blocks' SRP-PHAT of the observed signal for frequency range [250Hz 750Hz] in the complex plane

The proposed blind separation algorithm is used for separating the observed signal, and the SRP-PHAT of frequency range [250Hz, 750Hz] is used for voice activity detector to estimate the activity information of both speech sources. The frequency range [250Hz, 750Hz] had been chosen because the acoustical energy of speech in this frequency range is significant for voice processing [14]. There, Fig. 3 depicts all blocks' SRP-PHAT of observed signals for frequency range [250Hz, 750Hz] in the complex plane. Fig. 4 depicts time domain plots of the two outputs of the proposed separation algorithm. The two outputs are speech signals extracted for two speakers from the observed signal.

International Science Index, Computer and Information Engineering Vol:10, No:8, 2016 waset.org/Publication/10005265

Thus, Fig. 4 shows that the proposed algorithm can separate the two speech signals from the observed mixture. Informal listening tests suggest the good hearing quality of signal outputs from the proposed algorithm.

Also, the second-order blind signal separation (BSS) algorithm is used for separating the observed signal. This second-order BSS algorithm was proposed in [15] for the blind source separation. The outputs show a little separation level, and the separation did not have a good result in double-talk situation, see Table I.

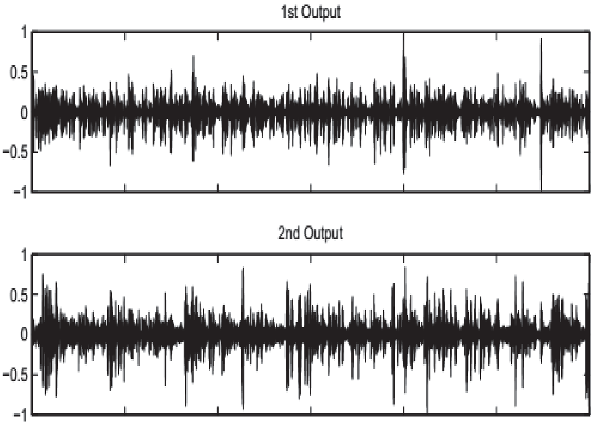

Fig. 4 Time domain plots of the proposed algorithm outputs

To quantify the performance of the proposed algorithm, the interference suppression (IS) and source distortion (SD) measures as in [16] are employed. As such, the speech signal from one speaker is viewed as the desired signal and other speech signals are interference. Table I shows the interference suppression and source distortion levels for the two outputs of the proposed beamformer and the second-order BSS algorithm. The table shows an improvement in the IS and SD levels of the proposed algorithm when compared with the second-order BSS algorithm. Here, the proposed blind speech separation algorithm offers a good interference suppression level ∼7 dB whilst maintaining a low distortion level (−21∼−24 dB) for the desired source.

VI. Conclusion

In this paper, a new blind speech separation algorithm in the frequency domain is developed for the two-speaker environment. Since the position of the sources is unknown, a voice activity detector using the SRP-PHAT is proposed for estimating the activity information of two speakers in observed signals. Based on the obtained activity information, an optimal beamformer is designed for each speech source to extract the desired signal in each frequency bin. For the algorithm evaluation, a simulation with a double-talk situation had been conducted by using real speech recordings. Simulation results show that the proposed algorithm manages to achieve a good noise suppression level ~7dB in a real double-talk environment whilst maintaining a low distortion level for each speech source.

Acknowledgment

This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under grant number C2014-26-01.

References

[1] K. Nakadai, K. Nakamura, and G. Ince, "Real-time super-resolution sound source localization for robots," IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 694–699, Oct. 2012.

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

[3] M. Fallon and S. Godsill, "Acoustic source localization and tracking of a time-varying number of speakers," IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 4, pp. 1409–1415, May 2012.

[4] H. Q. Dam, S. Nordholm, H. H. Dam, and S. Y. Low, "Postfiltering using multichannel spectral estimation in multi-speaker environments," EURASIP Journal on Advances in Signal Processing, pp. 1–10, Jan. 2008, ID 860360.

[5] N. Grbic´, X. J. Tao, S. Nordholm, and I. Claesson, "Blind signal separation using overcomplete subband representation," IEEE Transactions on Speech and Audio Processing, vol. 9, no. 5, pp. 524–533, July 2001.

[6] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 3, pp. 516–527, March 2011.

[7] J. Benesty, S. Makino, and J. Chen, Eds., Speech Enhancement, Springer-Verlag, 2005.

[8] P. Krishnamoorthy and S. R. Mahadeva Prasanna, "Two speaker speech separation by lp residual weighting and harmonics enhancement," International Journal of Speech Technology, vol. 13, no. 3, pp. 117–139, Sep. 2010.

[9] H. Q. Dam, "Blind multi-channel speech separation using spatial estimation in two-speaker environments," Journal of Science and Technology, Special Issue on Theories and Application of Computer Science, vol. 48, no. 4, pp. 109–119, Dec. 2010.

[10] H. Q. Dam and S. Nordholm, "Sound source localization for subband-based two speech separation in room environment," International Conference on Control, Automation and Information Sciences (ICCAIS), pp. 223– 227, Dec. 2013.

[11] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," IEEE Trans. on Speech and Audio Processing, vol. 11, no. 2, pp. 109–116, Mar. 2003.

[12] Shahab Faiz Minhas and Patrick Gaydecki, "A hybrid algorithm for blind source separation of a convolutive mixture of three speech sources," EURASIP Journal on Advances in Signal Processing, vol. 1, no. 92, pp. 1–15, Jan. 2014.

[13] M. Cobos, A. Marti, and J. J. Lopez, "A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling," IEEE Signal Processing Letters, vol. 18, no. 1, pp. 71–74, Nov. 2010.

[14] L. Saul, D. Lee, C. Isbell, Y LeCun, "Real time voice processing with audiovisual feedback: Toward autonomous agents with perfect pitch", Advances in Neural Information Processing Systems 15 (NIPS 2002), pp. 1205-1212, 2002.

[15] A Belouchrani, K Abed-Meraim, J-F Cardoso, E Moulines, "A Blind Source Separation Technique Using Second-Order Statistics", IEEE Transactions on Signal Processing, vol. 45, no. 2, pp. 434-444, Feb. 1997

[16] H. Q. Dam, S. Nordholm, H. H. Dam, and S. Y. Low, "Adaptive beamformer for hands-free communication system in noisy environments," IEEE Int. Symposium on Circuits and Systems, vol. 2, pp. 856–859, May 2005