A BSS-BASED APPROACH FOR LOCALIZATION OF SIMULTANEOUS SPEAKERS IN REVERBERANT CONDITIONS

Hamid Reza Abutalebi [1,2], *Hedieh Heli* [1], *Danil Korchagin* [2], *and Hervé Bourlard* [2]

[1] Speech Processing Research Lab (SPRL), Elec. and Comp. Eng. Dept., Yazd University, Yazd, Iran
[2] Idiap Research Institute, Martigny, Switzerland
phone: + (98) 351 8122396, fax: + (98) 351 8200144, email: habutalebi@yazduni.ac.ir
web: http://ee.yazduni.ac.ir/sprl, http://www.idiap.ch

ABSTRACT

In this paper, we address the localization of simultaneous speakers by means of Blind Source Separation (BSS) based algorithms. Considering BSS demixing filters as some blind null beamformer and producing an acoustical map from them, source localization can be achieved by identifying the local minima of this acoustical map. To improve the performance of this method in reverberant environments, we have proposed to replace the demixing filter with one, corresponding to the direct path only. This is done by keeping only the largest coefficient in each demixing filter and neglecting the other coefficients. Besides, the proposed method reduces the computational complexity. To further improve the computational efficiency of the localization method, we have also proposed the limitation of the frequency range within averaging procedure. The experimental results demonstrate improved accuracy and efficiency of the proposed method in the localization of multiple simultaneous sound sources in reverberant environments.

1. INTRODUCTION

Automatic speaker localization is an important task in several applications such as acoustic scene analysis, hands-free video conferences, hearing aids, and speech enhancement. It is also prerequisite for other processes like steering beamformer or pointing camera towards the sound sources.

In this paper, we concentrate on the category of multiple-speaker (sound source) localization methods that are based on Blind Source Separation (BSS). Generally these methods are divided in two groups: In the first group, by considering the relations between BSS problem and blind adaptive Multiple-Input-Multiple-Output (MIMO) system identification, BSS demixing filters are employed to estimate Time Difference Of Arrival (TDOA) in reverberant environments [1]; this can also be extended to the multi-dimensional case as well [2].

The second group of the BSS-based localization methods is based on the interpretation of BSS algorithms as a set of blind adaptive beamformers. To steer a beamformer toward a desirable source, the information of source location is necessary, while BSS algorithm recovers the original signals without any explicit information about source positions; This means that BSS demixing filters contain some useful and important information for localization of sound sources. In this group of methods, a directivity pattern is introduced based on the BSS demixing filters; then, the beam pattern is used to extract Direction Of Arrival (DOA). Some of the primary methods in this category used to extract location information of each source in each frequency bin separately. However, this causes a permutation problem specific to narrowband BSS [3, 4] which should be treated someway. Alternatively, [5] proposed an averaging procedure over all frequency bins and BSS outputs. The so-called BSS Averaged Directivity Pattern (BSS-ADP) provides useful information from a large range of frequencies, including the higher frequency regions which are potentially corrupted by spatial aliasing.

In this paper we concentrate on BSS-ADP and propose some modifications for improving the performance in (highly) reverberant environments and reducing computational complexity.

The rest of this paper is organized as follows. We review BSS algorithms and BSS-ADP method in Section 2. The proposed method is explained in Section 3. Section 4 explains the experimental evaluations. Finally, some concluding remarks are presented in Section 5.

2. SPEAKER LOCALIZATION USING THE BSS-ADP

2.1 Separation of convolutive mixtures

Fig. 1 shows the general BSS setup. In real-life environments, due to the reverberation, source signals are filtered by a MIMO mixing system, **H**, defined as:

$$\mathbf{H} = \begin{bmatrix} \boldsymbol{h}_{11} & \cdots & \boldsymbol{h}_{1P} \\ \vdots & \vdots & \vdots \\ \boldsymbol{h}_{Q1} & \cdots & \boldsymbol{h}_{QP} \end{bmatrix}, \quad (1)$$

which contains the Finite Impulse Response (FIR) filters $\boldsymbol{h}_{qp}(q = 1,2,...,Q,\ p = 1,2,...,P)$. Suppose that Q source signals $s_q(q = 1...Q)$ are mixed and observed at P sensors as:

$$x_p(n) = \sum_{q=1}^{Q} \sum_{k=0}^{M-1} \boldsymbol{h}_{qp,k} s_q(n-k), \quad (2)$$

where $\boldsymbol{h}_{qp,k}$ represents the k-th coefficient of the FIR filter from q-th source to p-th sensor (of length M). We assume that $Q \le P$.

The demixing (separation) system, **W**, is defined as:

$$\mathbf{W} = \begin{bmatrix} \boldsymbol{w}_{11} & \cdots & \boldsymbol{w}_{1Q} \\ \vdots & \vdots & \vdots \\ \boldsymbol{w}_{P1} & \cdots & \boldsymbol{w}_{PQ} \end{bmatrix}, \quad (3)$$

consisting a set of FIR filters $\boldsymbol{w}_{pq}$ (of length L) that produce Q separated signals:

$$y_q(n) = \sum_{p=1}^{P} \sum_{k=0}^{L-1} \boldsymbol{w}_{pq,k} x_p(n-k), \quad (4)$$

where $\boldsymbol{w}_{pq,k}$ represents the k-th coefficient of the demixing filter between p-th sensor to q-th output and it should be obtained blindly, i.e. without knowing $s_q(n)$ or $\boldsymbol{h}_{qp}$. To separate the source signals, s_q, without any information about the mixing system **H**, BSS algorithms force the output signals, y_q $(q = 1, 2, ..., Q)$, to be statistically independent. This is done by suitably adapting the weights of BSS demixing system **W**.

The general form of ideal separating filter matrix is shown in the frequency domain as follows [5]:

$$\mathbf{W}_{ideal}(f) = \text{Adj}\{\mathbf{H}(f)\} \cdot \mathbf{\Lambda} \cdot \mathbf{P}, \quad (5)$$

where Adj{.} operator computes the adjoint of a squared matrix. Matrix **P** shows the permutation and diagonal matrix **Λ** describes scaling of BSS outputs. The perfect separation will be achieved if the BSS demixing system converges to (5). This can be justified by considering the fact that in case of convergence, the overall mixing-demixing system would be reduced to a diagonal matrix:

$$\begin{aligned} \mathbf{C}_{ideal}(f) &= \mathbf{H}(f) \cdot \mathbf{W}_{ideal}(f) \\ &= \mathbf{H}(f) \cdot \text{Adj}\{\mathbf{H}(f)\} \cdot \mathbf{\Lambda} \cdot \mathbf{P} \\ &= \det\{\mathbf{H}(f)\} \cdot \mathbf{\Lambda} \cdot \mathbf{P}, \end{aligned} \quad (6)$$

where det{.} computes the determinant of a square matrix.

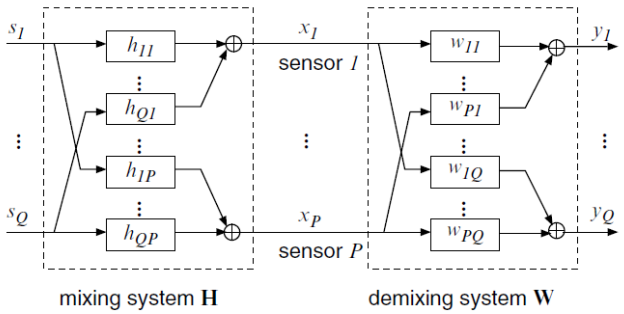


Figure 1 – BSS setup [6]

2.2 DOA extraction using averaged BSS directivity patterns

In this sub-section, we explain the BSS-ADP method proposed in [5] for a linear array of microphones and far-field assumption.

Convolutive blind source separation and adaptive beamforming have similarities in concept and structure. Both attempt to extract selected source signals from observed sensor mixtures by a filter array. Due to this simililarity, BSS demixing filters can also be interpreted as a set of blind adaptive null beamformers [7].

Considering Fig. 1 and applying Eq. (4) in frequency domain, the resultant output signals are obtained as

$$Y_q(f) = \sum_{p=1}^{P} \boldsymbol{w}_{pq}(f) X_p(f); \quad (7)$$

Then, at each frequency bin the directivity pattern for q-th output can be calculated using $\boldsymbol{w}_{pq}(f)$ [4]. The directivity pattern of each BSS output is defined as the magnitude squared response of a Multiple-Input-Single-Output (MISO) system of filters to plane waves coming from all possible directions, and is given by:

$$B_q(W, \theta, f) = \left| \sum_{p=1}^{P} \boldsymbol{w}_{pq}(f) e^{-j2\pi f d_p \sin(\theta)/c} \right|^2, \quad (8)$$

where $q = 1, 2, ..., Q$, θ is the direction of plane waves, c is the sound velocity, and d_p is the distance from p-th sensor to the reference sensor of the assumed linear array. This equation shows that the q-th directivity pattern, $B_q(W, \theta, f)$, is produced to extract the q-th source signal. Actually, these directivity patterns represent Q null beamformers, each of them creates Q-1 spatial nulls in the direction of Q-1 undesirable sources.

When the BSS null beamformers are considered in each frequency bin and each output, separately, the permutation problem is encountered. In order to solve this problem, [5] proposed to apply an averaging procedure before extracting the source locations. It consists of summing the BSS directivity patterns over the frequencies and over the P-1 best BSS outputs, as:

$$q^*(\theta, f) = \arg\max_{q} B_q(W, \theta, f), \quad (9)$$

$$\overline{B}(W, \theta) = \frac{1}{C} \int_{f_{\min}}^{f_{\max}} \sum_{\substack{q=1 \\ q \neq q^*(\theta,f)}}^{P} B_q(W, \theta, f)\, df, \quad (10)$$

where C is an arbitrary constant. In practice, integral is replaced with a summation over a finite number of frequency points from $f_{\min}$ to $f_{\max}$.

It is known that in high frequencies, spatial aliasing may occur. This will result in some errors in individual beam patterns. However, in averaging procedure of directivity patterns [5], only true spatial nulls add up coherently when summing over all BSS outputs and all frequencies; this partially decreases the effect of spatial aliasing at high frequencies. So the averaging procedure lets to collect useful location information from a large range of frequencies, even including those high frequency regions that are potentially corrupted by spatial aliasing in large microphone arrays. Source localization can then be achieved by identifying local minima in averaged directivity pattern diagram.

3. PROPOSED METHOD

In this section we propose some modifications to improve the performance of BSS-ADP method in reverberant environments and to reduce the computational complexity.

The definition (8) for BSS directivity pattern ignores the presence of reflection paths. As a result, the directivity pattern does not completely show the behaviour of the BSS algorithms under reverberant environments.

The idea behind the proposed modification is the considering direct path of sound propagation. According to the dominance of the direct propagation path in the acoustic impulse response, they can be very useful for source localization. In other words, the direct propagation paths deliver some meaningful location information. In reverberant environment, demixing filters would contain several coefficients with medium to large amplitude, where the largest one is considered as the direct path.

For instance, consider the case with 2 sources and 2 microphones. If there is no reverberation in the environment, i.e., under free field assumption, the single path filters of the mixing system **H** in frequency domain are in the form of:

$$\boldsymbol{h}_{qp}(f) = e^{-j2\pi f \tau_{qp}(\theta_q)}, \qquad (11)$$

where $\tau_{qp}(\theta_q) = d_p \sin(\theta_q)/c$ is the TDOA between the *p*-th sensor and the reference sensor for the *q*-th source with DOA θ_q and d_p is the distance from *p*-th sensor to the reference sensor of the linear array. Clearly, this filter is represented by only one coefficient in time domain, which is corresponding to the direct path between the *q*-th source and *p*-th sensor. Under real-life conditions, due to the reflection paths (reverberation), the mixing filters contain more than one coefficient in time domain. Nevertheless, the largest one (i.e. with the largest amplitude) corresponds to the direct path between the source and the sensor which is important in source localization.

Moreover, expanding (5) for the case of $P = 2$ sources and $\boldsymbol{\Lambda} = \mathbf{P} = \mathbf{I}$, we get:

$$\mathbf{W}_{ideal}(f) = \begin{bmatrix} \boldsymbol{h}_{22}(f) & -\boldsymbol{h}_{12}(f) \\ -\boldsymbol{h}_{21}(f) & \boldsymbol{h}_{11}(f) \end{bmatrix}, \qquad (12)$$

Therefore the ideal separation solution allows us to identify the filters of the acoustical demixing system. Considering (12) in time domain, we have:

$$\begin{aligned} \boldsymbol{w}_{11}(n) &= \boldsymbol{h}_{22}(n), \quad \boldsymbol{w}_{12}(n) = -\boldsymbol{h}_{12}(n) \\ \boldsymbol{w}_{21}(n) &= -\boldsymbol{h}_{21}(n), \quad \boldsymbol{w}_{22}(n) = \boldsymbol{h}_{11}(n) \end{aligned}, \qquad (13)$$

This equation clearly shows the correspondence between the largest coefficients in mixing and demixing filters, which is in turn corresponds to the direct path. Here, we propose to neglect all but only one (the largest) coefficient of the demixing filter and compute the directivity pattern (Eq. (8)) based on only the largest coefficient. For simplicity, hereafter, we refer to the proposed (modified) method as BSS-ADP Direct Path (or, briefly, BSS-ADP-DP).

It is noted that similar idea has been already applied in the problem of TDOA estimation [1], however, we use it here to compute directivity patterns of BSS outputs.

Furthermore, reducing the demixing filters to those with only one coefficient decreases the computational complexity drastically.

Similar to what mentioned in Section 2.2, the BSS-ADP-DP employs an averaging over the frequency range of $f_{\min}$ to $f_{\max}$. To improve the computational efficiency of the algorithm, we also propose the limitation of the frequency averaging range by reduction of $f_{\max}$. In turn, this may decrease the effect of spatial aliasing and improve the localization accuracy (as it is shown in Section 4).

4. EXPERIMENTAL EVALUATIONS

To demonstrate the performance of the BSS-ADP-DP in a reverberant situation, we have compared its localization accuracy with that of baseline BSS-ADP method (by considering the whole filter coefficients).

In our simulations, we have considered a linear microphone array that contains two omnidirectional microphones as shown in Fig. 2. The maximum length of microphone array is 21 cm and the sources are placed on a circle of 2 meters far from the center of array. Room Impulse Responses (RIRs) were simulated by an implementation of the Image method [8] at the sampling frequency $f_s = 16kHz$. Microphone signals were then generated by convolving source signals with the computed RIRs. The length of BSS filters is L = 1024 samples. Also, the frequency range for averaging procedure is $f_{\min} = 100\ Hz$ to $f_{\max} = 8000\ Hz$.

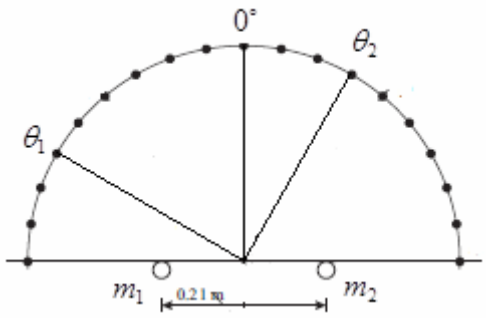
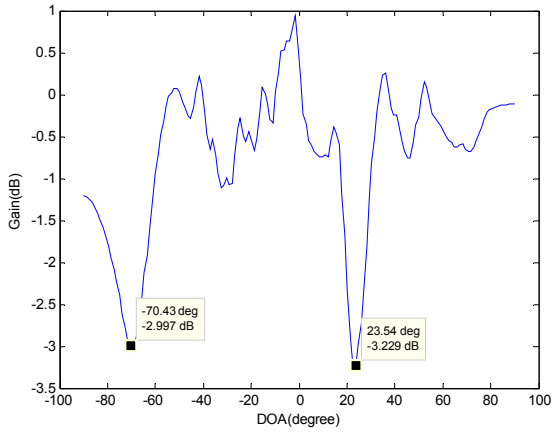


Figure 2 – microphone array and sources setup

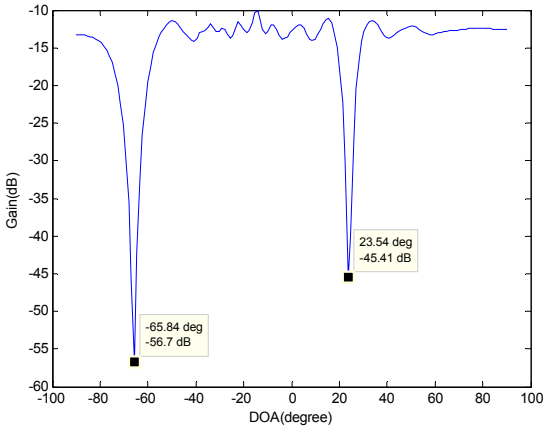
In the experiments, we assumed the case with two sources and two microphones. At first, we considered two microphones at the positions shown in Fig. 2 and two sources with DOA of $(\theta_1, \theta_2) = (-65, 25)$. The experiment was done under a highly reverberant situation, where the reverberation time was equal to $T_{60} = 1$ s.

Fig. (3-a) shows the directivity pattern for the baseline BSS-ADP. As shown, the method is not able to present a pattern with reliable nulls at the desired angles $(\theta_1, \theta_2) = (-65, 25)$.

Then the localization process was done via the proposed BSS-ADP-DP method. The resulted directivity pattern has been shown in Fig. (3-b). As shown, there are two distinct nulls within $\pm 2^\circ$ from exact DOAs.

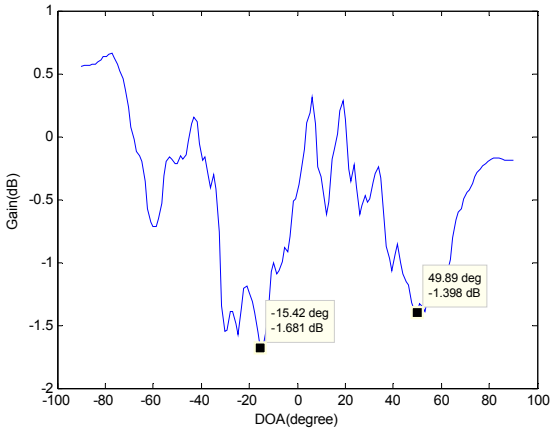


(a)

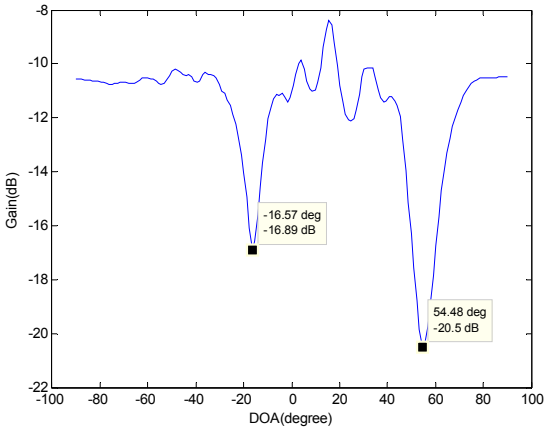


(b)

Figure 3- Directivity pattern of the (a) BSS-ADP, (b) BSS-ADP-DP, for the first experiment in a simulated room with $T_{60} = 1.0\ s$



(a)



(b)

Figure 4- Directivity pattern of the (a) BSS-ADP, (b) BSS-ADP-DP, for the second experiment in a simulated room with $T_{60} = 1.0\ s$

In the second experiment, we kept two microphones at the same positions (as shown in Fig. 2), while consider two other sources at DOAs of $(\theta_1, \theta_2) = (-20, 55)$. The outputs of the second experiment (the directivity patterns of BSS-ADP and BSS-ADP-DP) have been shown in (4-a) and (4-b). Obviously, the proposed method outperforms the baseline BSS-ADP in localization of two sources.

To compare the performance of the modified and baseline methods in a more objective manner, we repeated the localization procedure for 5000 times with different (random) (θ_1, θ_2) (each angle in the range of $[-90^\circ, +90^\circ]$). In this experiment, no minimum angle was kept between two sources, so the sources may fall very close. These tests were done for five different reverberant conditions ($T_{60} = 0.2, 0.5, 0.7, 0.9, 1.0$ s). In each case, the DOA error was computed as the absolute value of the difference of the resulted DOA and the exact one, summed over two sources. The results were averaged over 5000 cases. Table 1 shows average DOA error values of two mentioned methods in different reverberant conditions and for $f_{\max} = 8000Hz$. Clearly, the proposed modification has resulted in a superior method for the localization of multiple simultaneous sound sources in highly reverberant environments.

To examine the effect of lower $f_{\max}$, we repeated the tests on BSS-ADP-DP considering $f_{\max} = 5000Hz$ and $f_{\max} = 6000Hz$ (instead of $f_{\max} = 8000Hz$). As expected, our (informal) measurements showed the reduction of processing runtime. Also, Table 2 shows average DOA error values for the case of different $f_{\max}$ values. As shown, despite the reduction of frequency range within averaging procedure, the localization error remains in the same order (and even decreases in some cases, like the case of $f_{\max} = 6000Hz$). This can be justified by the reduction of the potential spatial aliasing.

Table 1- Average DOA error values for BSS-ADP and BSS-ADP-DP methods (for $f_{\max} = 8000Hz$)

	BSS-ADP	**BSS-ADP-DP**
$T_{60} = 0.2\ s$	2.02°	4.74°
$T_{60} = 0.5\ s$	6.82°	7.38°
$T_{60} = 0.7\ s$	10.61°	7.65°
$T_{60} = 0.9\ s$	15.35°	8.28°
$T_{60} = 1.0\ s$	16.96°	8.42°

Table 2- Average DOA error values for BSS-ADP-DP in different $f_{\max}$ values

	$f_{\max} =$ 5000 *Hz*	$f_{\max} =$ 6000 *Hz*	$f_{\max} =$ 8000 *Hz*
$T_{60} = 0.2\ s$	3.77°	3.89°	4.74°
$T_{60} = 0.5\ s$	6.88°	6.17°	7.38°
$T_{60} = 0.7\ s$	7.95°	7.04°	7.65°
$T_{60} = 0.9\ s$	8.37°	7.48°	8.28°
$T_{60} = 1.0\ s$	8.75°	7.95°	8.42°

5. CONCLUSION

In this paper, we used the ability of BSS algorithms to blindly identify the MIMO acoustical system. Using the proposed method, we could compute new BSS demixing filters which are formed by only the largest coefficient (corresponding to direct propagation path). Then, we used these filters to compute the averaged directivity pattern of [5]. We also limit the frequency range of the averaging to improve the computational efficiency of the proposed method (called BSS-ADP-DP). The method was evaluated for the case of two microphones and two sensors. The simulation results show the accuracy improvement of the method in multiple-speaker localization. In the future, we plan to extend the BSS-ADP-DP to localizing more sources and to nonlinear microphone arrays.

REFERENCES

[1] H. Buchner, R. Aichner, J. Stenglein, H. Teutsch, and W. Kellermann, "Simultaneous localization of multiple sound sources using blind adaptive MIMO filtering," in *Proc. ICASSP*, Philadelphia, PA, USA, 2005.

[2] A. Lombard, H. Buchner, and W. Kellermann, "Multidimensional localization of multiple sound sources using blind adaptive MIMO system identification," in *Proc. MFI*, Heidelberg, Germany, 2006.

[3] M. Ikram, D. Morgan, "A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation," in *Proc. ICASSP*, 2002, vol. 1, pp. 881–884.

[4] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda and F. Itakura , "Evaluation of blind signal separation method using directivity pattern under reverberant environments," in *Proc. ICASSP*, Istanbul, Turkey, 2000, vol. 5, pp. 3140–3143.

[5] A. Lombard, T. Rosenkranz, H. Buchner, and W. Kellermann, "Exploiting the self-steering capability of blind source separation to localize two or more sound sources in adverse environments," in *Proc. ITG Conference on Speech Communication*, Aachen, Germany, 2008.

[6] H. Buchner, R. Aichner and W. Kellermann, "TRINICON: a versatile framework for multichannel blind signal processing," in *Proc. ICASSP*, Montreal, Canada, 2004, vol. 3, pp. 889-892.

[7] S. Araki, S. Makino, Y. Hinamoto, R. Mukai, T. Nishikawa and H. Saruwatari, "Equivalence between frequency-domain blind source separation and frequency-domain adaptive beamforming for convolutive mixtures," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 11, pp. 1157–1166, 2003.

[8] E. Habets, Room impulse response generator. [online]. available: http://home.tiscali.nl/ehabets/rir_generator.html