Adaptive blind source separation with HRTFs beamforming preprocessing

Mounira Maazaoui, Karim Abed-Meraim and Yves Grenier

***Abstract*—We propose an adaptive blind source separation algorithm in the context of robot audition using a microphone array. Our algorithm presents two steps: a fixed beamforming step to reduce the reverberation and the background noise and a source separation step. In the fixed beamforming preprocessing, we build the beamforming filters using the Head Related Transfer Functions (HRTFs) which allows us to take into consideration the effect of the robot's head on the near acoustic field. In the source separation step, we use a separation algorithm based on the l_1 norm minimization. We evaluate the performance of the proposed algorithm in a total adaptive way with real data and varying number of sources and show good separation and source number estimation results.**

***Index Terms*—Adaptive blind source separation, fixed beamforming, head related transfer functions**

I. Introduction

Blind source separation (BSS) [1] is the ability to estimate the source signals using their mixtures, without any prior knowledge of the mixing process or the looked up sources. In this article, we investigate blind source separation in a real environment for the application of robot audition. Robot audition consists in the aptitude of an humanoid to understand its acoustic environment, separate and localize sources, identify speakers and recognize their emotions. This complex task is one of the target points of the ROMEO project[1] that we work on. This project aims to build an humanoid (ROMEO) that can act as a comprehensive assistant for persons suffering from loss of autonomy. Our task in this project is focused on the blind source separation topic using a microphone array (more than 2 sensors).

One of the main challenges of blind source separation remains to obtain good BSS performance in a real reverberant environment. To reduce the reverberation of a room, a beamforming preprocessing can be a solution [2]. A fixed beamforming, contrarily to an adaptive one, does not depend on the sensors data, the beamformer is built for a set of fixed desired directions. In [3], we proposed a two-stage iterative blind source separation technique where a fixed beamforming is used in a preprocessing step. The advantage of the fixed beamforming is that the beamforming filters are generally estimated offline, using the microphone array geometry and the acoustic field clues. To overcome the problem of the array geometry modeling and take into account the influence of the robot's head on the received signals, we use the Head Related Transfer Functions (HRTFs) of the robot's head as steering vectors to build the fixed beamformer [3].

[1]Romeo project: www.projetromeo.com

In the robot audition context, the number of sources are unknown and can change dynamically. In this paper, we propose a fully adaptive blind source separation algorithm that can deal with the dynamic change of the number of sources. The main contributions of this article are: 1) the *adaptive* blind source separation algorithm with a fixed beamforming preprocessing using HRTFs and 2) the *adaptive* estimation of the number of sources that changes dynamically thanks to the fixed beamforming preprocessing.

II. A two step separation algorithm

Assume we are in a real room with N sound sources $\mathbf{s}(t) = [s_1(t), \ldots, s_N(t)]^T$ and an array of M microphones with outputs denoted by $\mathbf{x}(t) = [x_1(t), \ldots, x_M(t)]^T$, where t is the time index. We assume that we are in an overdetermined case with $M > N$. As we are in a real environment context, the output signals in the time domain are modeled as the sum of the convolution between the sound sources and the impulse responses of the different propagation paths between the sources and the sensors, truncated at the length of $L+1$:

$$\mathbf{x}(t) = \sum_{l=0}^{L} \mathbf{h}(l)\, \mathbf{s}(t-l) + \mathbf{n}(t) \tag{1}$$

where $\mathbf{h}(l)$ is the l^{th} impulse response matrix and $\mathbf{n}(t)$ is a noise vector that will be neglected in the rest of the article[2].

In the frequency domain, the output signals at the time-frequency bin (f,k) can be approximated as: $\mathbf{X}(f,k) \simeq \mathbf{H}(f)\,\mathbf{S}(f,k)$, where $\mathbf{X}(f,k) = [X_1(f,k), \ldots, X_M(f,k)]^H$ (respectively $\mathbf{S}(f,k) = [S_1(f,k), \ldots, S_N(f,k)]^H$) is the Short-time Fourier transform (STFT) of $\{\mathbf{x}(t)\}_{1\leq t\leq T}$ (respectively $\{\mathbf{s}(t)\}_{1\leq t\leq T}$) in the frequency bin $f \in \left[1, \frac{N_f}{2}+1\right]$ and the time bin $k \in [1, N_T]$, and $\mathbf{H}$ is the Fourier transform of the mixing filters $\{\mathbf{h}(l)\}_{0\leq l\leq L}$. Using an appropriate separation criterion, our objective is to find for each frequency bin a separation matrix $\mathbf{F}(f)$ that leads to an estimation of the original sources in the time-frequency domain:

$$\mathbf{Y}(f,k) = \mathbf{F}(f)\,\mathbf{X}(f,k) \tag{2}$$

The inverse short time Fourier transform of the estimated sources in the frequency domain $\mathbf{Y}$ allows the recovery of the estimated sources $\mathbf{y}(t) = [y_1(t), \ldots, y_N(t)]^T$ in the time domain. Separating the sources for each frequency bin introduces the permutation problem which is solved by the

[2]We consider discrete sound sources and the diffuse background noise energy is supposed to be negligible comparing to the source ones.

method described in [4] based on the signals correlation between two adjacent frequencies.

The separation matrix $\mathbf{F}(f)$ is estimated using a two-step blind separation algorithm:

1) Fixed beamforming preprocessing step: the signals in the sensors are filtered using the offline estimated beamforming filters $\mathbf{B}(f)$, the output signal is $\mathbf{Z}(f,k) = \mathbf{B}(f)\mathbf{X}(f,k)$. More details about this step are presented in the next section.
2) Source separation step: we apply a blind source separation algorithm to the outputs of the beamformer. We use a sparsity separation criterion based on the l_1 norm minimization to estimate the separation matrix $\mathbf{W}(f)$ [3]. The optimization technique used to update the separation matrix $\mathbf{W}(f)$ is the natural gradient proposed by Amari et *al.* in 1996 [5], the update equation is written as:

$$\mathbf{W}_{j+1}(f) = \mathbf{W}_j(f) - \mu \nabla\psi(\mathbf{W}_j(f))\,\mathbf{W}_j^H(f)\,\mathbf{W}_j(f) \quad (3)$$

$\psi(\mathbf{W}(f))$ is our loss function, μ is an adaptation step and j refers to the frame number. The output signal is $\mathbf{Y}(f,k) = \mathbf{W}(f)\mathbf{Z}(f,k)$ and this separation algorithm will be referred to as BSS-l_1.

The final separation matrix $\mathbf{F}(f)$ is written as the combination of the results of those two steps: $\mathbf{F}(f) = \mathbf{W}(f)\mathbf{B}(f)$

III. Beamforming preprocessing

A. Offline estimation of the beamforming filters

In the case of robot audition, the geometry of the microphone array is fixed once for all. To build the fixed beamformers, we need to determine the "desired" steering directions and the characteristics of the beam pattern. The beamformers are estimated only once for all scenarii using these spatial information and independently of the measured mixture in the sensors.

In the robot audition context, the microphones are often fixed in the head of the robot and it is hard to model the microphone array manifold in this case. In fact, the phase and magnitude response models of the free field steering vectors [3] model do not take into account the influence of the head on the surrounding acoustic fields. So we propose to use the Head Related Transfer Functions[4] (HRTFs) as steering vectors $\{\mathbf{a}(f,\theta)\}_{\theta\in\Theta}$, where $\Theta = \{\theta_1,\ldots,\theta_{N_S}\}$ is a group of N_S *a priori* chosen steering directions [3]. The HRTF takes into account the head and microphone array geometry and the influence of the head on the near acoustic field. We extend the notion of HRTFs to a microphone array case. Let $h_m(f,\theta)$ be the HRTF at frequency f from the emission point located at θ to the m^{th} sensor. The steering vector is then $\mathbf{a}(f,\theta) = [h_1(f,\theta),\ldots,h_M(f,\theta)]^T$. Given the equation of the steering vector, one can estimate the beamformer filters that will achieve the desired beam pattern according to the desired direction response θ_i using the least-square technique [2]:

$$\mathbf{b}(f,\theta_i) = \frac{\mathbf{R}_{\mathbf{aa}}^{-1}(f)\,\mathbf{a}(f,\theta_i)}{\mathbf{a}^H(f,\theta_i)\,\mathbf{R}_{\mathbf{aa}}^{-1}(f)\,\mathbf{a}(f,\theta_i)} \quad (4)$$

where $\mathbf{R}_{\mathbf{aa}}(f) = \frac{1}{N_S}\sum_{\theta\in\Theta}\mathbf{a}(f,\theta)\,\mathbf{a}^H(f,\theta)$. Given K desired steering directions $\theta_1,\ldots,\theta_K$, the beamforming matrix is $\mathbf{B}(f) = [\mathbf{b}(f,\theta_1),\ldots,\mathbf{b}(f,\theta_K)]^T$.

B. Beamforming filtering

In our case, we fix K steering directions such as the corresponding beams cover all the useful space directions. We consider $\{\mathbf{B}(f)\}_{1\le f\le\frac{N_f}{2}+1}$ a set of fixed beamforming filters of size $K\times M$, $K \ge N$. Those filters are calculated offline, before the beginning of the processing, for each frequency, as shown in the previous subsection. The outputs of the beamformers at each frequency f are: $\mathbf{Z}(f,k) = \mathbf{B}(f)\mathbf{X}(f,k)$.

C. Highest energy beams selection and source number estimation

After the beamforming, the signal is spatially filtered toward the K chosen steering directions $\theta_1,\ldots,\theta_K$. The beams who are the closest to the sources capture the most of their energy. From this observation, we propose to estimate the number of sources by selecting the beams that contain the highest energy. This can be done as follow (this processing is going to be referred to as *BeamSelect*) :

1) In each frequency bin f, after the beamforming filtering, we select the N_{max} steering directions corresponding to the N_{max} beams that give the highest energies.
2) We build over all the selected steering direction a histogram that corresponds to their overall number of occurrence as shown in figure 1.
3) After a proper thresholding, we select the peaks that corresponds to the highest number of selected beams over all the frequencies. The filters corresponding to those beams are our final beamforming filters $\tilde{\mathbf{B}}(f)$, the number of peaks is an estimation of the number of sources and the corresponding steering directions provide us with a rough estimation of the directions of arrival (DOA).

IV. Adaptive blind source separation algorithm with fixed beamforming preprocessing

In this section, we present the implementation details of our two step separation algorithm in a fully adaptive context with a varying number of sources. The main difficulty in this case is to adapt the separation matrix $\mathbf{F}(f) = \tilde{\mathbf{B}}(f)\mathbf{W}(f)$ from one frame to the next one. The idea is to update $\tilde{\mathbf{B}}(f)$ and $\mathbf{W}(f)$ separately. As the number of source can vary, $\tilde{\mathbf{B}}(f)$ is updated in each frame by selecting the beams with the highest energies, and thus, the number of sources and the direction of arrivals are also estimated. The separation matrix $\mathbf{W}(f)$ of the frame $j-1$ is used as initialization matrix for the BSS-l_1 algorithm in the frame j. But as the number of sources

[3]The steering vectors represent the phase delays of a plane wave evaluated at the microphone array elements.

[4]The HRTFs characterize how the signal emitted from a specific direction is received at a sensor fixed in a head and are generally used in a binaural context.

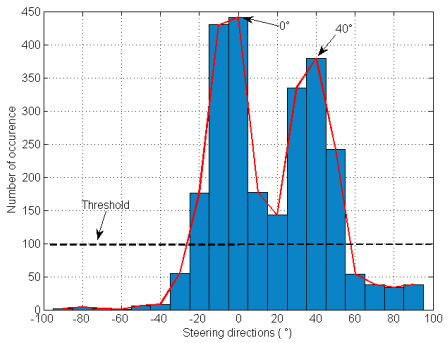

Figure 1: Estimation of the source number using fixed beamforming

from the frame $j-1$ to the frame j can be different, a size adjustment of $\mathbf{W}(f)$ is necessary. In the following the details of our algorithm.

Initialization: Frame1:

1) Fixed beamforming preprocessing:
   a) beamforming filtering: $\mathbf{Z}_1(f,:) = \mathbf{B}(f)\,\mathbf{X}_1(f,:)$
   b) $\left[\tilde{\mathbf{Z}}_1(f,:), N_1, \mathbf{doa}_1\right] = \text{BeamSelect}(\mathbf{Z}_1(f,:))$
2) $[\mathbf{Y}_1(f,:), \mathbf{W}_1(f)] = \text{BSS-}l_1\left(\tilde{\mathbf{Z}}_1(f,:), \mathbf{W}_0\right)$

Frame j:

1) Fixed beamforming preprocessing:
   a) beamforming filtering: $\mathbf{Z}_j(f,:) = \mathbf{B}(f)\,\mathbf{X}_j(f,:)$
   b) $\left[\tilde{\mathbf{Z}}_j(f,:), N_j, \mathbf{doa}_j\right] = \text{BeamSelect}(\mathbf{Z}_j(f,:))$
2) Source separation depending on the number of estimated sources N_j
   a) if $N_j = 1$, $\mathbf{Y}_j(f,:) = \tilde{\mathbf{Z}}_j(f,:)$
   b) if $N_j = N_{j-1}$, $[\mathbf{Y}_j(f,:), \mathbf{W}_j(f)] = \text{BSS-}l_1\left(\tilde{\mathbf{Z}}_j(f,:), \mathbf{W}_{j-1}\right)$
   c) if $N_j > N_{j-1}$,
      i) Estimate the index $\mathbf{ind}$ of the new sources using the estimated DOAs $\mathbf{doa}_{j-1}$ and $\mathbf{doa}_j$
      ii) Modify the separation matrix $\mathbf{W}_{j-1}(f)$ by adding columns and rows in the corresponding new sources index $\mathbf{ind}$
      iii) $[\mathbf{Y}_j(f,:), \mathbf{W}_j(f)] = BSS - l_1\left(\tilde{\mathbf{Z}}_{j-1}(f,:), \mathbf{W}_{j-1}\right)$
   d) if $N_j < N_{j-1}$,
      i) Estimate the index $\mathbf{ind}$ of the vanished sources using the estimated DOAs $\mathbf{doa}_{j-1}$ and $\mathbf{doa}_j$
      ii) Modify the separation matrix $\mathbf{W}_{j-1}(f)$ by deleting the columns and rows of the corresponding vanished source index $\mathbf{ind}$
      iii) $[\mathbf{Y}_j(f,:), \mathbf{W}_j(f)] = \text{BSS-}l_1\left(\tilde{\mathbf{Z}}_{j-1}(f,:), \mathbf{W}_{j-1}\right)$

V. Experimental results

A. Experimental database

To evaluate the proposed BSS techniques, we built two databases: a HRTFs database and a speech database. We recorded the HRTF database in the anechoic room of Telecom ParisTech. As we are in a robot audition context, we model the future robot by a child size dummy ($1m20$) for the sound acquisition process, with 16 sensors fixed in its head (*cf.* figure 2). We measured 504 HRTF for each microphone: 72 azimuth angles from 0° to 355° with a 5° step, 7 elevation angles. The HRTF database is available for download[5].

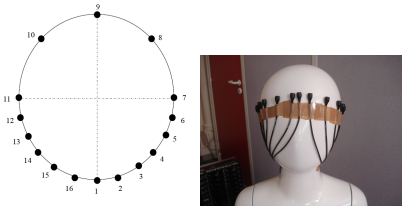
Figure 2: The detailed configuration of the microphone array

The test signals were recorded in a moderately reverberant room where the reverberation time is $\text{RT}_{30} = 300$ ms. We chose to evaluate the proposed algorithm on a separation of 2 sources: the first source is always the one placed at 0° and the second source is chosen from 30° to 90°. The distance between the sources and the microphone array is $1m20$. The output signals $\mathbf{x}(t)$ are the convolutions of 20 pairs of 15s of speech sources (male and female speaking French and English) by two of the impulse responses $\{\mathbf{h}(l)\}_{0\leq l\leq L}$ measured for the cited directions of arrival. The signals are sampled at 16KHz, the length of the adaptive analysis window is 1s, the length of the shift and the STFT window is 64ms and the step size of the optimization algorithm is $\mu = 0.05$.

B. Results and discussion

First, we want to show the effect of the beamforming preprocessing by evaluating the Signal-to-Interference Ratio [6] of the separated sources a) after the beamforming filtering, the inter-beams angle in 5° (BF[5°]) b) with the blind source separation only (BSS-l_1) c) with the beamforming preprocessing without beams selection (BF[5°]+BSS-l_1) and d) with the beamforming preprocessing and the highest energy beams selection (BF[5°]+BS+BSS-l_1). Figure 3 shows that the beamforming preprocessing BF[5°]+BSS-l_1 improves the SIR of the estimated sources comparing to the use of the blind source separation algorithm only BSS-l_1. Besides, the beamforming preprocessing with the selection of the beams with the highest energy (BF[5°]+BS+BSS-l_1) gives the best separation results.

We now vary the number of sources between one and two. We estimate the number of sources using our method (BF), and two eigenvalues based methods (EIG1 [7] and EIG2 based on a simple thresholding of the sorted eigenvalues of the covariance matrices in the frequency domain [8]). Figure 4 and 5 show the average of the estimated number of sources for 20 pairs of speakers in each of the shown DOA. The results of the source number estimation of our method are close to EIG2 ones. But our method is a direct result from

[5]http://www.tsi.telecom-paristech.fr/aao/?p=347

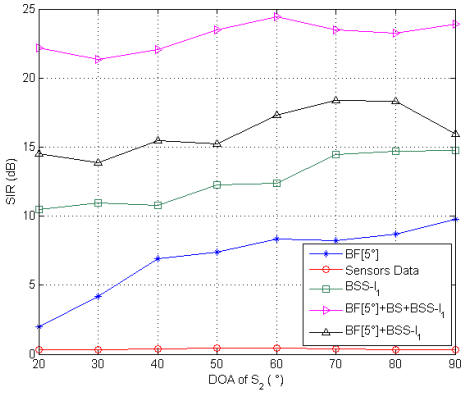

Figure 3: SIR comparison in a real environment: source 1 is at 0° and source 2 varies from 20° to 90° with a step of 10°

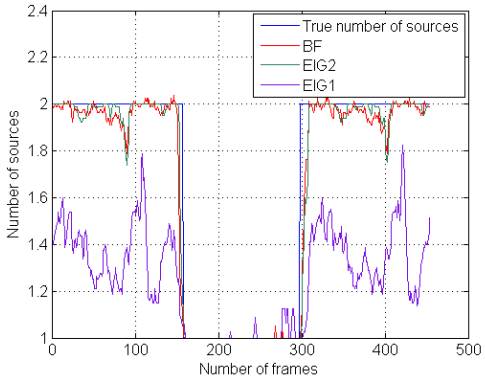

Figure 4: The number of sources estimated through the temporal frames

the beamforming preprocessing, it is simple to implement and does not need any calculation other than the peaks estimation. EIG2 takes much more calculation time than our method due to the calculation of the covariance matrices and the singular values decomposition.

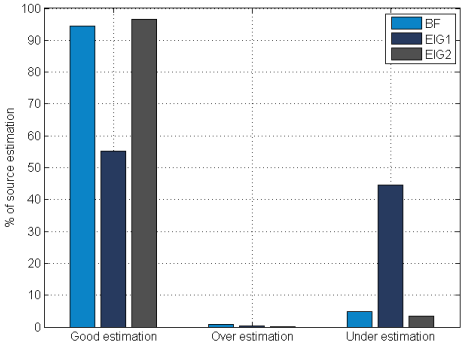

Figure 5: Results of the estimation of the number of sources over all the frames

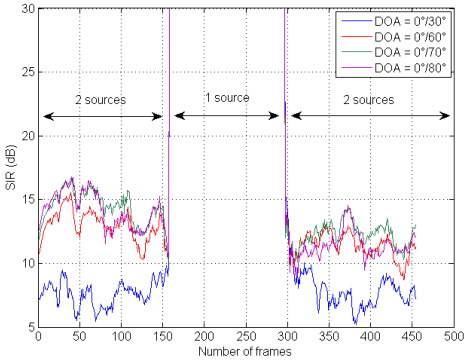

Figure 6: SIR of the separated sources for a number of sources varying between 1 and 2 and for different DOA

Figure 6 shows the average SIR of all the pairs of mixtures for different direction of arrivals. Our algorithm follows the dynamic change of the number of sources and converge quickly. We recall that the separation matrix is initialized once and that the adaptation is totally automatic and depends on the number of estimated sources.

VI. Conclusion

We propose a complete adaptive blind source separation algorithm for robot audition context. Our system can estimate the number of sources and separate them thanks to its two steps separation process: the first step is a beamforming preprocessing which allows us to reduce the reverberation effect and estimate the number of sources, the second step is a source separation step based on the l_1 norm minimization. Our estimation of the number of sources is simple, not time consuming and suitable for a real time application of this algorithm, which is going to be our next work.

References

[1] Pierre Comon and Christian Jutten, *Handbook of Blind Source Separation , Independent Component Analysis and Applications*, Elsevier, 2010.

[2] Jacob Benesty, Jingdong Chen, and Yiteng Huang, *Microphone Array Signal Processing, Chapter 3: Conventional beamforming techniques*, Springer, 1rst edition, 2008.

[3] Mounira Maazaoui, Yves Grenier, and Karim Abed-Meraim, "Blind source separation for robot audition using fixed beamforming with hrtfs," *21th Annual Conference on the International Speech Communication Association, Interspeech 2011*, 2011.

[4] Wang Weihua and Huang Fenggang, "Improved method for solving permutation problem of frequency domain blind source separation," *6th IEEE International Conference on Industrial Informatics*, pp. 703–706, July 2008.

[5] S. Amari, A. Cichocki, and H. H. Yang, "A new learning algorithm for blind signal separation," *Advances in Neural Information Processing Systems*, pp. 757–763, 1996.

[6] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 1462 –1469, July 2006.

[7] Jingqing Luo and Zhiguo Zhang, "Using eigenvalue grads methods to estimate the number of source," *International Conference on Software Process, ICSP*, 2000.

[8] K. Yamamoto, F. Asano, van W.F.G. Rooijen, E.Y.L. Ling, T. Yamada, and N. Kitawaki, "Estimation of the number of sound sources using support vector machines and its application to sound source separation," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, April 2003.