FROM BINAURAL TO MULTIMICROPHONE BLIND SOURCE SEPARATION USING FIXED BEAMFORMING WITH HRTFS

Mounira Maazaoui, Yves Grenier and Karim Abed-Meraim

Institut TELECOM, TELECOM ParisTech, CNRS-LTCI 37/39, rue Dareau, 75014, Paris, France
maazaoui@telecom-paristech.fr, yves.grenier@telecom-paristech.fr, abed@telecom-paristech.fr

ABSTRACT

In this article, we are interested in the problem of blind source separation (BSS) for the robot audition, we study the performance of blind source separation with a varying number of sensors in a microphone array placed in the head of an infant size dummy. We propose a two stage blind source separation algorithm based on a fixed beamforming preprocessing using the head related transfer functions (HRTF) of the dummy and a separation algorithm using a sparsity criterion. We show that in the case of robot audition, the use of a multisensor array improves significantly the performance of the source separation algorithm, as compared to the binaural case, up to a limit number of microphones studied in this paper.

Index Terms— Blind source separation, beamforming, binaural BSS, multisensors BSS, robot audition

1. INTRODUCTION

Blind source separation (BSS) consists in recovering the original sources from their mixtures, using the received microphone signals and without prior knowledge of the mixing process. Our work is focused on BSS using a microphone array for robot audition, in the context of the ROMEO project[1]. 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 which aims to build a humanoid (ROMEO) to help aged people in their everyday lives.

Blind source separation has been tackled many times [1]. In this article, we study the influence of the number of the microphones on the BSS performance in a robot audition context. For that, we have considered a BSS approach that combines spatial information (source location, beamforming) with structural information of the source signals (statistical independence, sparsity, etc ...). Indeed, it is shown in [2] and recently in [3] that such an approach leads to an improved BSS performance. In this paper, we propose to use a fixed beamforming technique based on the HRTFs knowledge as a preprocessing step, followed by a separation technique based on source sparsity in the time-frequency domain.

With respect to this separation method, a main objective consists in studying the effect of the number of microphones on the source separation quality. More specifically, we attempt to find out the "optimal" number of microphones that should be used for robot audition given a specified array geometry (*i.e.* the microphones are around the robot's head as shown in figure 4, in the 16 sensors case). This is done by comparing the BSS performance for different array sizes (from 2 to 16 microphones) in order to find the number of microphones above which the performance gain becomes negligible.

This work is funded by the Ile-de-France region, the General Directorate for Competitiveness, Industry and Services (DGCIS) and the City of Paris, as a part of the ROMEO project

[1]Romeo project: www.projetromeo.com

2. SIGNAL MODEL

Assume N sound sources $\mathbf{s}(t) = [s_1(t), \ldots, s_N(t)]^T$ and an array of M microphones. The outputs are denoted by $\mathbf{x}(t) = [x_1(t), \ldots, x_M(t)]^T$, where t is the time index and $M \geq N$. When the sources are placed in a real reverberant environment, 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 coefficient and $\mathbf{n}(t)$ is a noise vector. In the frequency domain, when the analysis window of the Short Time Fourier Transform (STFT) is longer than the length of the mixing filter, the output signals at the time-frequency bin (f,k) can be approximated as:

$$\mathbf{X}(f,k) \simeq \mathbf{H}(f)\, \mathbf{S}(f,k) \tag{2}$$

where $\mathbf{X}$ (respectively $\mathbf{S}$) is the STFT of $\{\mathbf{x}(t)\}_{1\leq t\leq T}$ (respectively $\{\mathbf{s}(t)\}_{1\leq t\leq T}$) and $\mathbf{H}$ is the Fourier transform of the mixing filters $\{\mathbf{h}(l)\}_{0\leq l\leq L}$. In the blind source separation task, our goal is to find for each frequency bin a separation matrix $\mathbf{F}(f)$ that leads to an estimation of the original sources:

$$\mathbf{Y}(f,k) = \mathbf{F}(f)\, \mathbf{X}(f,k) \tag{3}$$

This introduces the well known permutation problem: from one frequency to the adjacent one, the order of the estimated sources may be different. The permutation problem can be solved by the method described in [4] based on the signal correlation between two adjacent frequencies. The sources in the time domain can be recovered by taking the inverse short time Fourier transform of the estimated sources in the frequency domain, after solving the permutation problem.

3. COMBINED BEAMFORMING AND BSS ALGORITHM

To compare the performance of the binaural case and the multimicrophone one in the context of robot audition, we present here the two steps blind source separation algorithm based on the sparsity of

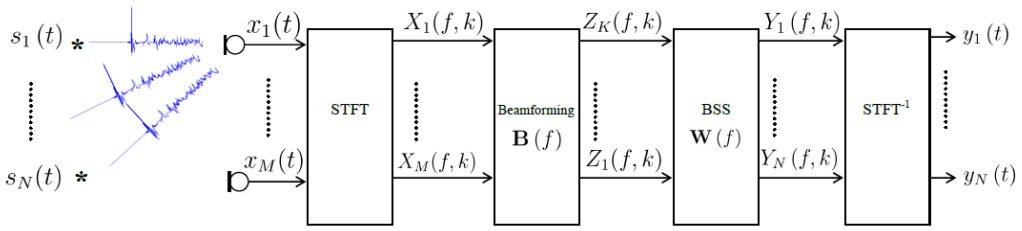

Fig. 1: The processing scheme of the combined beamforming-BSS algorithm

the sources in the time-frequency domain and using a fixed beamforming preprocessing. Figure 1 shows the separation process.

3.1. Fixed beamforming using HRTFs

The role of the beamforming is essentially to reduce the reverberation so that equation (2) is satisfied with better precision, which leads to an improvement in the BSS quality. The beamformer also reduces the interferences coming from directions other than the looked up ones.

The desired steering direction is determined by a localization technique or arbitrarily. Estimating the direction of arrivals of the sources as it was done in [2] is time consuming and not always accurate in the reverberant environments. As an alternative solution, we propose to build a fixed beamformer containing K fixed beams with desired directions chosen in such a way that they cover all useful directions. We consider $\{\mathbf{B}(f)\}_{1\leq f\leq N_f/2}$ a set of fixed beamforming filters of size $K \times M$, where N_f is the length of the Fourier analysis window and K is the number of the desired beams, $K \geq N$. The outputs of the beamformers at each frequency f are:

$$\mathbf{Z}(f,k) = \mathbf{B}(f)\,\mathbf{X}(f,k) \tag{4}$$

To design a fixed beamformer that will achieve the desired beam pattern (according to a desired direction response), the least-square (LS) technique is used [5] and thus the steering vectors are needed. In the case of robot audition, the microphones are often fixed in the head of the robot and it is generally hard to know exactly the geometry of the microphone array (*cf.* figure 3). Besides, the phase and magnitude of the steering vectors 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 (HRTFs) as steering vectors $\{\mathbf{a}(f,\theta)\}_{\theta\in\Theta}$, where $\Theta = \{\theta_1,\ldots,\theta_K\}$ is a group of K *a priori* chosen steering directions (*cf.* figure 2). The HRTF characterizes how the signal emitted from a specific direction is received at a sensor fixed in a head. It takes into account the geometry of the head, and thus the geometry of the microphone array.

We use the following steering vector:

$$\mathbf{a}(f,\theta) = [h_1(f,\theta),\ldots,h_M(f,\theta)]^T \tag{5}$$

where $h_m(f,\theta)$ is the HRTF at frequency f from the emission point located at θ to the m^{th} sensor. Given equation (5), one can express the normalized LS beamformer for a desired direction θ_i as [5]:

$$\mathbf{b}(f,\theta_i) = \frac{\mathbf{R}_{\mathbf{aa}}^{-1}(f)\,\mathbf{a}(f,\theta_i)}{\mathbf{a}^H(f,\theta_i)\,\mathbf{R}_{\mathbf{aa}}^{-1}(f)\,\mathbf{a}(f,\theta_i)} \tag{6}$$

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 $\mathbf{B}(f)$ is:

$$\mathbf{B}(f) = [\mathbf{b}(f,\theta_1),\ldots,\mathbf{b}(f,\theta_K)]^H \tag{7}$$

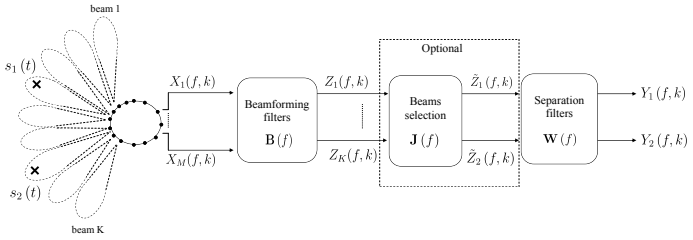

Fig. 2: Beamforming with fixed DOAs: the selection of the beams with the highest energy is optional

Beam selection

After the fixed beamforming, the signal is spatially filtered toward the K chosen steering directions $\theta_1,\ldots,\theta_K$. As an alternative solution, we propose to choose the N beamformer outputs with the highest energies that correspond to the closest beams to the sources, supposing that the energies of the sources are quite close (*cf.* figure 2). This step leads to a reduced computational cost in the following BSS step.

3.2. Blind source separation

The blind source separation step consists in estimating a separation matrix $\mathbf{W}(f)$ that leads to separated sources at each frequency bin f. The separation matrix is estimated from the beamformers outputs $\mathbf{Z}(f,k)$, the estimated sources are then written as:

$$\mathbf{Y}(f,k) = \mathbf{W}(f)\,\mathbf{J}(f)\mathbf{Z}(f,k) \tag{8}$$

where $\mathbf{J}(f)$ is a $N\times K$ selection matrix that selects the N "highest energy" beam outputs.

The separation matrix $\mathbf{W}(f)$ is estimated using a sparsity criterion. We assume that the sources in the time-frequency domain are the sparsest state to reach from their mixtures and we use the l_1 norm minimization criterion:

$$\min_{\mathbf{W}}\sum_{i=1}^{N}\sum_{k=1}^{N_T}|Y_i(f,k)| \tag{9}$$

where $Y_i(f,k)$ is the $(f,k)^{th}$ bin of the i^{th} extracted signal. The update given by equation 10 of $\mathbf{W}(f)$ using the natural gradient descent technique [6]:

$$\mathbf{W}_{t+1}(f) = \mathbf{W}_t(f) - \mu\nabla\psi(\mathbf{W}_t(f))\,\mathbf{W}_t^H(f)\mathbf{W}_t(f) \tag{10}$$

where $\psi(\mathbf{W}(f)) = \sum_{i=1}^{N}\sum_{k=1}^{N_T}|Y_i(f,k)|$ is the cost function, $\nabla\psi(\mathbf{W}(f))$ is the gradient of $\psi(\mathbf{W}(f))$ and t refers to the

iteration index (or time index for an adaptive processing). The final separation matrix $\mathbf{F}(f)$ is written as:

$$\mathbf{F}(f) = \mathbf{W}(f)\mathbf{J}(f)\mathbf{B}(f) \tag{11}$$

Once the sources are separated, the permutation problem is solved as in [4].

4. EXPERIMENTAL RESULTS

4.1. Experimental database

To evaluate the proposed BSS techniques and the effect of the number of sensors on the separation performance, we built two databases: a HRTF database and a speech database. We recorded the HRTF database in the anechoic room of Telecom ParisTech (*cf.* figure 3). 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 3). We measured 504 HRTF for each microphone as follow:

- 72 azimuth angles from 0° to 355° with a 5° step
- 7 elevation angles: -40°, -27°, 0°, 20°, 45°, 60° and 90°

The HRTFs were measured by a Golay codes process [7] at a sampling frequency of 48 kHz downsampled to 16 kHz. The HRTF database is available for download at `http://www.tsi.telecom-paristech.fr/aao/?p=347`.

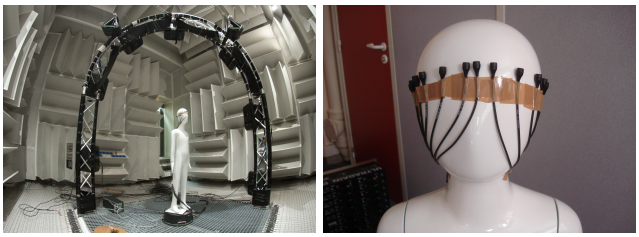

Fig. 3: The dummy in the anechoic room (left) and the microphone array of 16 sensors (right)

We also recorded, with the same dummy, a reverberant speech database to evaluate and compare the proposed methods. The output signals $\mathbf{x}(t)$ are the convolutions of 2 sources (male and female speaking French and English) by the impulse responses $\{\mathbf{h}(l)\}_{0\leq l\leq L}$ measured from different angles of arrivals in a moderately reverberant room where the reverberation time is $\mathrm{RT}_{30} = 300$ ms. Source 1 is at 0° and source 2 varies from 20° to 90°. 30 different source pairs were used for each DOA for the evaluation of the separation performance with respect to the change of the number of sensors. The characteristics of the signals and the BSS algorithms are summarized in table 1.

Sampling frequency	16 kHz
Analysis window	Hanning
Analysis window length	2048
Shift length	1024
μ	0.2
Signals length	5s
Number of iterations	100

Table 1: Parameters of the blind source separation algorithms

4.2. Results and discussion

We evaluate the proposed two stage algorithm by the Signal-to-Interference Ratio (SIR), the Signal-to-Distortion Ratio (SDR) and the Signal-to-Artifact Ratio (SAR) calculated using the BSS-eval toolbox [8]. All the SIR, SDR and SAR curves are the averages of the results obtained by using the 30 separation cases. We used the two-stage algorithm with fixed beam pattern: we suppose that the sound sources come from the front of the dummy, so we consider fixed beams from -90 to 90, with a step of 5°. The number of sensors varies from 2 to 16 according to the distribution shown in figure 4.

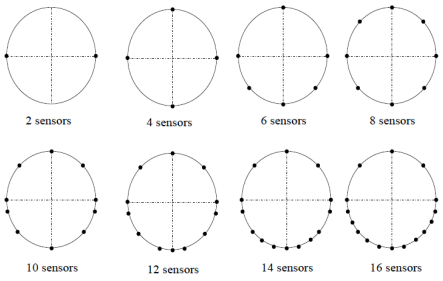

Fig. 4: A view from above of the sub-arrays configuration using 2 to 16 sensors

Figure 5 shows the increase of the SIR thanks to the beamforming preprocessing in the case of 16 sensors. *Sensors Data* is the SIR of the mixture, *BF[5°]* is the SIR when only the beamforming is used, *BSS-l_1* is the SIR when only the separation algorithm is used, *BF[5°]+BSS* the SIR of BSS with beamforming preprocessing without lobe selection and *BF[5°]+BS+BSS* is the SIR of the BSS with beamforming preprocessing using the lobe selection.

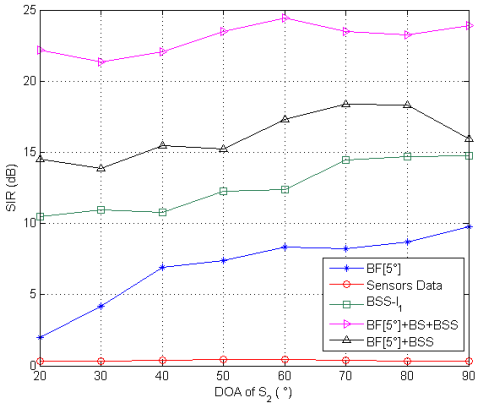

Fig. 5: SIR comparison in a real environment: source 1 is at 0° and source 2 varies from 20° to 90° - Effect of the beamforming preprocessing on the SIR of the estimated sources

Now if we vary the number of sensors from 2 to 16, we can see in figure 6 that, for any number of sensors, blind source separation using a beamforming preprocessing performs better that a source separation algorithm only.

Figure 7 shows the variation of the average SIR versus the number of microphones for a separation of 2 sources: source 1 is placed

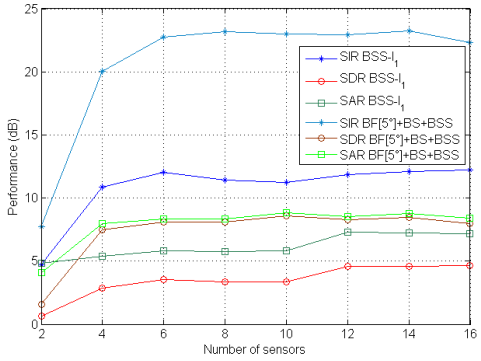

Fig. 6: SIR, SDR and SAR variation with the number of sensors for the separation of 2 sources from different DOAs: effect of the beamforming preprocessing

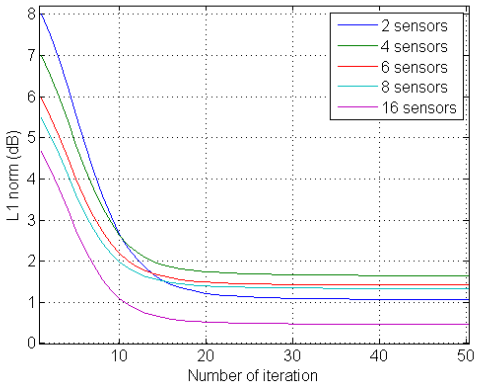

Fig. 8: The mean of the l_1 norm through the iteration of the BSS algorithm

at 0° and source 2 varies from 20° to 90°. We can see a significant performance improvement in terms of SIR when the number of sensors increases. When the number of microphones is $M \geq 8$, no significant gain is observed for all the DOAs. This observation rises our interest and can be explained by: first the moderately reverberant room, we expect more important effect of the increase of the number of microphone in more reverberant room, second the effect of the beamforming lobe shape. We also note that the bigger is the difference between the directions of arrivals, the faster is the convergence to a constant state of performance with respect to the number of microphones.

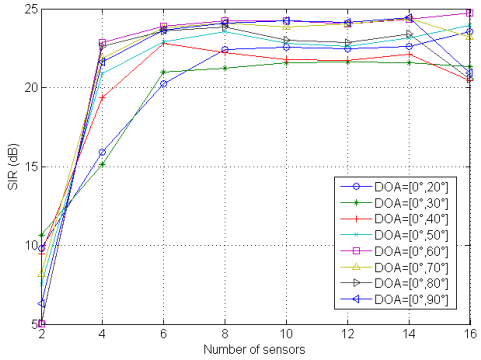

Fig. 7: SIR, SDR and SIR variation with the number of sensors for the separation of 2 sources and for different DOAs using the BF[5°]+BS+BSS algorithm

In figure 8, we compare the iterative algorithm convergence rate with respect to the number of sensors. This experiment shows that the former is roughly independent from the array size.

5. CONCLUSION

In this paper, we studied the effect of the number of microphones on the blind source separation performance in a robot audition context, which is useful for an effective choice of the microphone array size. We used a two-stage BSS algorithm: the first stage consists in a fixed beamforming preprocessing to reduce the reverberation and noise coming from directions other that the desired ones and the second stage is a BSS algorithm based on a sparsity criterion. In this context and using the proposed microphone array geometry, we show that when using a microphone array (in the range of [2,16] sensors in our case) the performance of the separation increases significantly with respect to the binaural case until $M = 8$ for all the tested direction of arrivals, then no significant gain is observed for $M \geq 8$.

6. REFERENCES

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

[2] Heping Ding Lin Wang and Fuliang Yin, "Combining superdirective beamforming and frequency-domain blind source separation for highly reverberant signals," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, 2010.

[3] Karim Abed-Meraim Mounira Maazaoui, Yves Grenier, "Blind source separation for robot audition using fixed beamforming with hrtfs," *21th Annual Conference on the International Speech Communication Association, Interspeech*, 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] Jacob Benesty, Jingdong Chen, and Yiteng Huang, *Microphone Array Signal Processing, Chapter 3: Conventional beamforming techniques*, Springer, 1rst edition, 2008.

[6] 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.

[7] S. Foster, "Impulse response measurement using golay codes," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '86*, Apr. 1986, vol. 11, pp. 929 – 932.

[8] 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.