Blind Source Separation for Robot Audition using Fixed Beamforming with HRTFs

Mounira Maazaoui, Yves Grenier, Karim Abed-Meraim

Blind Source Separation for Robot Audition 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

We present a two stage blind source separation (BSS) algorithm for robot audition. The algorithm is based on a beamforming preprocessing and a BSS algorithm using a sparsity separation criterion. Before the BSS step, we filter the sensors outputs by beamforming filters to reduce the reverberation and the environmental noise. As we are in a robot audition context, the manifold of the sensor array in this case is hard to model, so we use pre-measured Head Related Transfer Functions (HRTFs) to estimate the beamforming filters. In this article, we show the good performance of this method as compared to a single stage BSS only method.

Index Terms: blind source separation, beamforming, robot audition

1. Introduction

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]. This project aims to build an humanoid (ROMEO) to help aged people in their everyday lives. In this project, we focus on blind source separation (BSS) using a microphone array (more than 2 sensors). In a blind source separation task, the separation should be done from the received microphone signals without prior knowledge of the mixing process. The only knowledge is limited to the array geometry. Source separation is the most important step for human-robot interaction: it allows latter tasks like speakers identification, speech and motion recognition and environmental sound analysis.

Blind source separation problem has been tackled several times [2] and one of the main challenges remains to have good BSS performance in a high reverberant environments. One way to handle the reverberation problem is beamforming. Beamforming consists in estimating a spatial filter that operates on the outputs of a microphone array in order to form a beam with a desired directivity pattern [3]. It is useful for many purposes, particularly in enhancing a desired signal from its measurement corrupted by noise, competing sources and reverberation [3]. Beamforming can be fixed or adaptive. A fixed beamforming, contrarily to the adaptive one, does not depend on the sensors data, the beamformer is built for a set of fixed desired directions. In this article, we propose a two stage blind source separation technique where a fixed beamforming is used as a preprocessing. However, in a beamforming task, we need to know the manifold of the sensor array, which is sometimes hard to model for the robot audition case. 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 propose to use the Head Related Transfer Functions (HRTFs) of the robot's head to build the fixed beamformer.

Wang *et al.* propose to use a beamforming preprocessing where the steering directions are the direction of arrivals of the sources [4]. This suppose that the direction of arrival are known *a priori*. The authors evaluate their method in a determined case (2 and 4 sources) with a circular microphone array. Saruwatari *et al.* present a combined Independent Component Analysis (ICA) and beamforming method: first they perform a subband ICA and estimate the direction of arrivals (DOA) of the sources using the directivity patterns, second they use the estimated DOA to build a null beamforming, and third they integrate the subband ICA and the null beamforming by selecting the most suitable separation matrix in each frequency [5]. In this article, we propose to use a fixed beamforming preprocessing with fixed steering directions, independently from the direction of arrival of the sources, and we compare this preprocessing method to the Wang *et al.* one. We are interested in studying the effect of the beamforming as a preprocessing tool so we are not going to include the algorithm of [5] in our evaluation (the authors of [5] use the beamforming as a separation method alternatively with ICA). We present promising results using two different fixed beamformers as preprocessing and a sparsity based BSS algorithm.

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 of the sensors array are denoted by $\mathbf{x}(t) = [x_1(t), \ldots, x_M(t)]^T$, where t is the time index and $M \geq N$. In a general case, 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

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.

mixing filters $\{\mathbf{h}(l)\}_{0\leq l\leq L}$. 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 and scaling problems: from one frequency to the adjacent one, the order and the scale of the estimated sources may be different. The permutation problem can be solved by the method described in [6] based on the signals correlation between two adjacent frequencies. The scale problem is solved by the method proposed in [7]. 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

We present here a two step blind separation algorithm based on a fixed beamforming preprocessing (*cf.* figure 1).

3.1. Beamforming pre-processing

We consider $\{\mathbf{B}(f)\}_{1\leq f\leq \frac{N_f}{2}}$ 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, $K \geq N$. Those filters are calculated beforehand (*cf.* section 4). The outputs of the beamformers at each frequency f are:

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

The role on the beamformer is essentially to reduce the reverberation (and consequently, equation 2 is better satisfied leading to improved BSS quality) and the interferences coming from space directions other than the looked up ones.

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:

$$\mathbf{Y}(f,k) = \mathbf{W}(f)\,\mathbf{Z}(f,k) \tag{5}$$

The separation matrix $\mathbf{W}(f)$ is estimated using a sparsity criterion. We assume that the STFT of the sources are the sparsest state to reach and we use the l_1 norm minimization criterion:

$$\min_{\mathbf{W}} \sum_{i=1}^{N}\sum_{k=1}^{N_T} |Y_i(f,k)| \text{ such that } \|\mathbf{W}\| = 1 \tag{6}$$

where $Y_i(f,k)$ is the $(f,k)^{th}$ bin of the i^{th} source and $\|.\|$ is any matrix norm. The update equation of $\mathbf{W}(f)$ using the natural gradient descent technique [8] is:

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \mu\nabla\psi(\mathbf{W}_t)\,\mathbf{W}_t^T\mathbf{W}_t \tag{7}$$

where $\psi(\mathbf{W}) = \sum_{i=1}^{N}\sum_{k=1}^{N_T}|Y_i(f,k)|$, $\nabla\psi(\mathbf{W})$ is the gradient of $\psi(\mathbf{W})$ and t refers to the iteration (or time for an adaptive processing) index. The final separation matrix $\mathbf{F}$ is:

$$\mathbf{F}(f) = \mathbf{W}(f)\,\mathbf{B}(f) \tag{8}$$

4. Fixed beamforming with HRTFs

In the case of robot audition, the geometry of the microphone array is fixed once for all. Once the array geometry is fixed and the desired steering direction is determined (by a localization technique or arbitrarily), the fixed beamformer takes full advantage of these spatial information to design the desired beam pattern. Thus, the desired characteristics of the beam pattern (the beamwidth, the amplitude of the sidelobes and the position of nulls) are obtained for all scenarii and calculated only once.

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 [3]. In the case of robot audition, the microphone are often fixed in the head of the robot and it is generally hard to know exactly the manifold of the microphone array (*cf.* figure 5). Besides, the phase and magnitude response models 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 $\Theta = \{\theta_1,\ldots,\theta_{N_S}\}$ is a group of N_S *a priori* chosen steering directions (*cf.* figure 2). The HRTFs characterize how the signal emitted from a specific direction is received at a sensor fixed in a head. It takes into account the geometry and the manifold of the head, and thus of the microphone array. 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:

$$\mathbf{a}(f,\theta) = [h_1(f,\theta),\ldots,h_M(f,\theta)]^T \tag{9}$$

Given equation (9), one can express the normalized LS beamformer for a desired direction θ_i as [3]:

$$\mathbf{b}(f,\theta_i) = \frac{\mathbf{R}_{\mathbf{aa}}^{-1}(f)\,\mathbf{a}(f,\theta_i)}{\mathbf{a}^H(f,\theta_i)\,\mathbf{R}_{\mathbf{aa}}^{-1}(f)\,\mathbf{a}(f,\theta_i)} \tag{10}$$

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

$$\mathbf{B}(f) = [\mathbf{b}(f,\theta_1),\ldots,\mathbf{b}(f,\theta_K)]^T \tag{11}$$

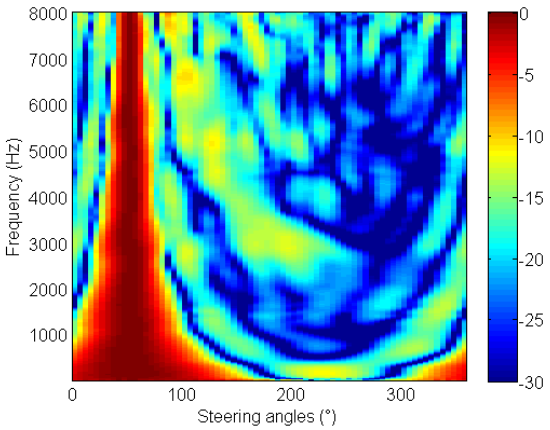

Figure 2: Example of a beam pattern using HRTFs for $\theta_i = 50°$ (in dB)

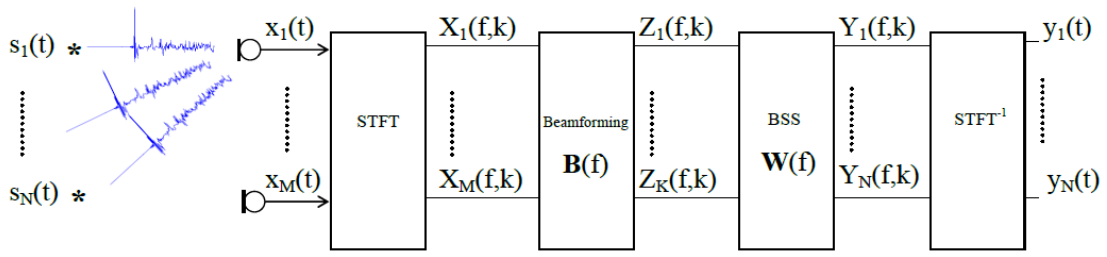

Figure 1: The processing scheme of the combined beamforming-BSS algorithm

4.1. Beamforming with known DOA

If the direction-of-arrivals (DOAs) of the sources are known *a priori*, mainly by a source localization method, the beamforming filters are estimated using this spatial information of the sources location (*cf.* figure 3). Therefore, the desired directions are the DOAs of the sources and we select the corresponding HRTFs to build the desired response vectors $\mathbf{a}(f,\theta)$.

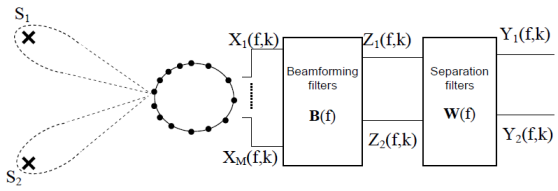

Figure 3: Beamforming with known DOAs

4.2. Beamforming with fixed DOA

Estimating the DOAs of the sources is time consuming and not always accurate in the reverberant environments. As an alternative solution, we propose to build K fixed beams with arbitrary desired directions, and then chose the N beamformer outputs with the highest energy, corresponding to the beams that are the closest to the sources (we suppose that the energy of the sources are quite close). The K directions are chosen in such a way they cover all useful space directions (*cf.* figure 4).

In general, the sources may have a big difference in their energy levels and the source that has the highest energy could be detected more than once by different beams. So the selection of the outputs with the highest energy is done in such a way that they are not correlated, *i.e.* their coherence is below a certain threshold fixed *a priori*.

5. Experiments and results

5.1. 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 anechoïc room of Telecom ParisTech (*cf.* figure 5). 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 5). We measured 504 HRTF for each microphone as follow:

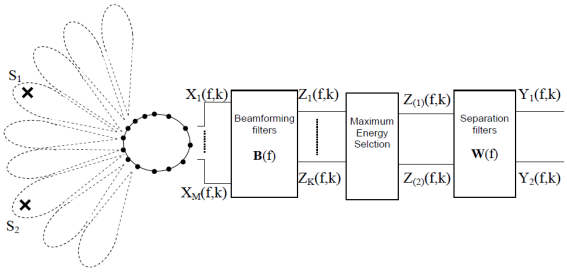

Figure 4: Beamforming with fixed steering directions

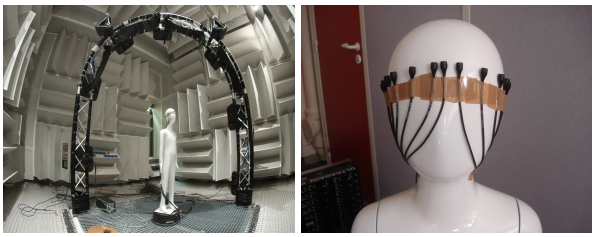

Figure 5: The dummy in the anechoïc room (left) and the microphone array of 16 sensors (right)

- 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 [9] at a sampling frequency of 48 KHz downsampled to 16 KHz. The HRTF database is available for download[1].

We also recorded, with the same dummy, a reverberant speech database to evaluate and compare the proposed methods. The test signals were recorded in a moderately reverberant room which reverberation time is $\mathrm{RT}_{30} = 300$ ms (*cf.* figure **??**). The output signals $\mathbf{x}(t)$ are the convolutions of 10 pairs of speech sources (male and female speaking french and English) by the impulse responses $\{\mathbf{h}(l)\}_{0\leq l\leq L}$ measured from two angles of arrivals. We used 3 different pairs of DOAs on the horizontal plan (the reference is the head of Theo where the elevation =0°): 0°/-30°, 0°/27° and 0°/90°.

The characteristics of the signals and the BSS algorithms are summarized in table 1.

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

Sampling frequency	16 KHz
Analysis window	Hanning
Analysis window length	2048
Shift length	1024
μ	0.2
Signals length	5s
Number of iterations	500

Table 1: Parameters of the blind source separation algorithms

5.2. Results and discussion

We evaluate the proposed two stage algorithm by the Signal-to-Interference Ratio (SIR) and the Signal-to-Distortion Ratio (SDR) calculated using the BSS-eval toolbox [10]. The following results are the mean SIR (*cf.* figure 6) and the mean SDR (*cf.* figure 7) of 10 pairs of source separation cases from different directions of arrivals. The algorithms that we evaluate in this section are:

- Sparsity based BSS algorithm (BSS)
- Two stage algorithm with know DOAs (K-DOA+ BSS)
- Two stage algorithm with fixed DOAs: 7 beams from -90 to 90 with a step of 30° (F-DOA[30°]+ BSS)
- Two stage algorithm with fixed DOAs: 13 beams from -90 to 90 with a step of 15° (F-DOA[15°]+ BSS)

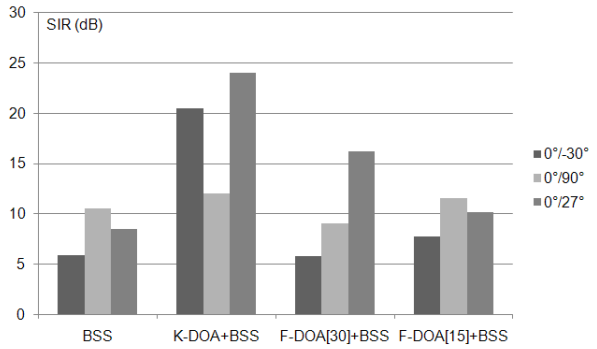

Figure 6: SIR comparison with different direction-of-arrivals (DOA)

The results show an improvement of the SIR when the beamforming preprocessing is used. The SDR is considerably improved in the case of known DOAs, as no distortions are introduced by a difference between the fixed steering direction and the real DOA. Some errors may occur in the selection of the beams with the highest energy in the case of the fixed DOAs beamforming, especially in the low frequencies and this may affect the SDR and SIR performance. But as the results show, we still have good performance with the fixed DOA beamforming preprocessing. We also have a gain in the processing time as we do not have to estimate the DOAs in each separation case.

6. Conclusion

In this article, we present a two stage blind source separation algorithm for robot audition. This algorithm is based on a fixed beamforming preprocessing, with known or fixed DOAs and a BSS algorithm exploiting the sparsity of the sources in the time-frequency domain. The beamforming preprocessing improves the separation performance as it reduces the reverberation and noise effects. The maximum gain is obtained when the sources DOAs are known. However, we propose also a beamforming preprocessing with fixed DOAs that has good performance and do not use an estimation of the DOAs, which represent a gain in the processing time.

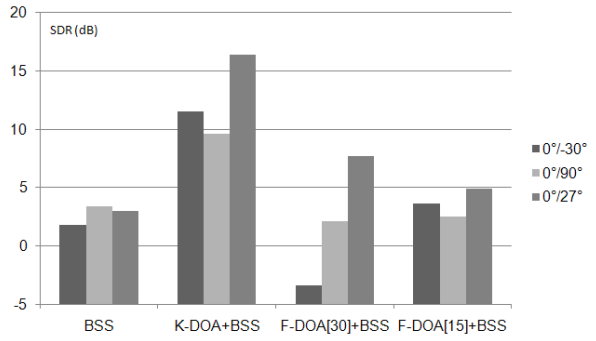

Figure 7: SDR comparison with different direction-of-arrivals (DOA)

7. References

[1] Romeo project: www.projetromeo.com.

[2] P. Comon and C. Jutten, *Handbook of Blind Source Separation , Independent Component Analysis and Applications*. Elsevier, 2010.

[3] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing, Chapter 3: Conventional beamforming techniques*, 1st ed. Springer, 2008.

[4] H. D. Lin Wang and F. Yin, "Combining superdirective beamforming and frequency-domain blind source separation for highly reverberant signals," in *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, 2010.

[5] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, and K. Shikano, "Blind source separation combining independent component analysis and beamforming," *EURASIP Journal on Applied Signal Processing*, pp. 1135–1146, 2003.

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

[7] S. Douglas and M. Gupta, "Scaled natural gradient algorithms for instantaneous and convolutive blind source separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP 2007, Honolulu, HI*, vol. 2, Apr. 2007, pp. II–637 –II–640.

[8] S. Amari, A. Cichocki, and H. H. Yang, *A New Learning Algorithm for Blind Signal Separation*. MIT Press, 1996, pp. 757–763.

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

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