FREQUENCY DOMAIN BLIND SOURCE SEPARATION FOR ROBOT AUDITION USING A PARAMETERIZED SPARSITY CRITERION

Mounira Maazaoui, Yves Grenier, Karim Abed-Meraim

Institut TELECOM, TELECOM ParisTech, CNRS-LTCI 37/39, rue Dareau,75014, Paris, France

ABSTRACT

In this paper, we introduce a modified l_p norm blind source separation criterion based on the source sparsity in the time-frequency domain. We study the effect of making the sparsity constraint harder through the optimization process, making the parameter p of the l_p norm vary from 1 to nearly 0 according to a sigmoid function. The sigmoid introduces a smooth l_p norm variation which avoids the divergence of the algorithm. We compared this algorithm to the regular l_1 norm minimization and an ICA based one and we obtained promising results.

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 motions. This complex task is one of the target points of the ROMEO project [6]. This project aims to build an humanoid (ROMEO) to help aged people in their everyday lives. Our task in this project is focused on the source separation topic using a microphone array (more than 2 sensors). 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. 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.

The problem of blind source separation has been tackled by many authors [4], and we present here some of the state-of-the-art methods related to robot audition. Tamai *et al.* performed sound source localization and separation in a real environment with delay and sum beamforming and frequency band selection using three rings microphone array with 32 microphones [9]. Yamamoto *et al.* proposed a robot audition system for automatic speech recognition of simultaneous speech where a source separation technique based on geometric constraints was used as a preprocessing for the speech recognition [12]. This system was implemented in the humanoids SIG2 and Honda ASIMO (with a 8 sensors microphone array), as a part of a more complete system for robot audition named HARK [5]. Saruwatari *et al.* proposed a two-stage blind source separation for a mixed binaural signals of an humanoid. They combined a single-input multiple-output model based on independent component analysis (ICA) and a binary mask processing [8].

Blind source separation can be done in the time domain or in the time-frequency domain. In the time domain, the source mixture is modeled as a convolutive mixture between the sources and the impulse responses of the different paths from the sources to the microphones. In this case, we have to find a separation filter according to the considered separation criterion. In the frequency domain, the convolutive mixture is approximated by an instantaneous one, and the problem becomes easier as we have to find a separation matrix instead of a separation filter. But this has to be done for each frequency bin which gives rise to the permutation and scale problem. Another advantage of the time-frequency domain is the sparsity of the signals in this domain. A signal is sparse when it is zero or nearly zero in most of its samples.

In this article, we propose a separation criterion based on the sparsity maximization of the estimated source, therefore the minimization of their l_p norm. We compare this algorithm to ICA and a $l_1$minimization using an experimental database for robot audition.

2. SIGNAL MODEL

Assume N sound sources $\mathbf{s}(t) = [s_1(t), \ldots, s_N(t)]^T$ and an array of M microphones with $M > N$. 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. 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 mixing filters $\{\mathbf{h}(l)\}_{0\leq l\leq L}$. Our goal is to use an appropriate criterion to find, for each frequency bin, a separation matrix $\mathbf{W}(f)$ that leads to an estimation of the original sources:

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

This introduces a permutation problem: from one frequency to the adjacent one, the order of the estimated sources may be different. This can be solved by the method described in [11] based on the signals correlation between two adjacent

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

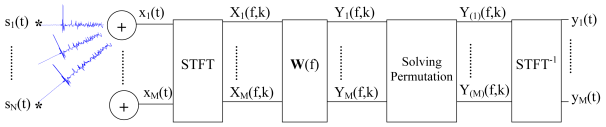

Figure 1: Blind source separation modules

frequencies. The sources in the time domain can be recovered by making the inverse short time Fourier transform of the estimated sources in the frequency domain, after solving the permutation problem.

3. PARAMETERIZED L_P NORM ALGORITHM

3.1 Principle

We assume that the time-frequency representation of the sources is the sparsest state to reach from the mixtures: we look for a separation matrix $\mathbf{W}$[1] that leads to the sparsest estimated sources $\mathbf{Y}$ in every frequency bin. To measure the sparsity of the signal $\mathbf{Y}$, we use the l_p norm[2], with $0 < p \leq 1$ and we propose the following loss function:

$$\psi(\mathbf{W}) = \sum_{i=1}^{N} \left| \sum_{k=1}^{N_T} |Y_i(k)|^p \right|^{\frac{1}{p}} \tag{4}$$

As a separation criterion, we minimize the l_p norm of the estimated sources with respect to the separation matrix $\mathbf{W}$, under a unit norm constraints for $\mathbf{W}$:

$$\min_{\mathbf{W}} \psi(\mathbf{W}(f)) \text{ such that } \|\mathbf{W}\| = 1 \tag{5}$$

where $\|.\|$ is any matrix norm. The l_1 norm is the most used sparsity measure thanks to its convexity. However, the closer the parameter p to 0, the harder is the sparsity measure: the extreme example is the $l_0(\mathbf{x})$ norm which is the number of the non-zero elements in the vector $\mathbf{x}$. We have no prior knowledge on the choice of the parameter p for the blind source separation task. As we want to reach the sparsest possible state of the estimated sources in the frequency domain, we propose to make the sparsity constraint harder through the iterations of the optimization process. The idea is to decrease p from the less hard sparsity constraint $p = 1$ to the hardest one $p \simeq 0$. But changing the l_p norm through the iterations of the algorithm may lead to a divergence, so we propose to decrease the parameter p according to a sigmoid curve with a very small step change, so the convergence of the algorithm may not be disturbed (*cf.* figure 2). The loss function is then:

$$\hat{\psi}(\mathbf{W}) = \sum_{i=1}^{N} \left| \sum_{k=1}^{N_T} |Y_i(k)|^{p(t)} \right|^{\frac{1}{p(t)}} \tag{6}$$

[1]From now on, we remove the frequency index f to make equations clearer : $\mathbf{W} = \mathbf{W}(f)$, $\mathbf{Y} = [\mathbf{Y}(f,k)]_{1 \leq k \leq N_T}$ of dimension $N \times N_T$ and $\mathbf{X} = [\mathbf{X}(f,k)]_{1 \leq k \leq N_T}$ of dimension $M \times N_T$ where N_T is the number of the temporal frames in the STFT.

[2]For $0 < p < 1$, l_p is a quasi-norm

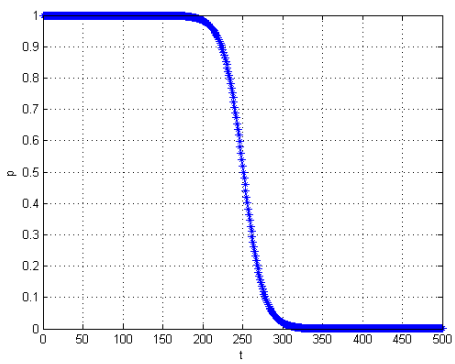

Figure 2: The parameter p as a logistic function, $p = p(t) = \frac{1}{1-exp(-L+\frac{(t-1)2L}{NbIter})}$, L is the range of computation of the sigmoid and $NbIter = 500$ is the iteration number

3.2 Proposed algorithm

To solve the constrained minimization of equation (5), we use the natural gradient optimization method with coefficient normalization. The natural gradient is a modified gradient search method proposed by Amari et *al.* in 1996 [1]. The standard gradient search direction is altered according to the local Riemannien structure of the parameter space. This guarantees the invariance of the natural gradient search direction to the statistical relationship between the parameters of the model and leads to a statistically efficient learning performance [2].

The gradient update of the separation matrix $\mathbf{W}$ is given by:

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

where

$$\tilde{\nabla} \hat{\psi}(\mathbf{W}) = \nabla \hat{\psi}(\mathbf{W}) \mathbf{W}^T \mathbf{W} \tag{8}$$

is the natural gradient of the function $\hat{\psi}(\mathbf{W})$ and t refers to the iteration (or time for an adaptive processing) index. From (7) and (8), and considering the unit norm constraint, the update of the separation matrix $\mathbf{W}$ is:

$$\begin{cases} \tilde{\mathbf{W}}_{t+1} = \mathbf{W}_t - \mu \nabla \hat{\psi}(\mathbf{W}_t) \mathbf{W}_t^T \mathbf{W}_t \\ \mathbf{W}_{t+1} = \frac{\tilde{\mathbf{W}}_{t+1}}{\|\tilde{\mathbf{W}}_{t+1}\|} \end{cases} \tag{9}$$

The differential of $\hat{\psi}(\mathbf{W})$ is:

$$\mathrm{d}\hat{\psi}(\mathbf{W}) = \mathbf{f}(\mathbf{Y}) \mathrm{d}\mathbf{Y}^T \tag{10}$$

where $\mathbf{f}(\mathbf{Y}) = p(t) |\mathbf{Y}|^{p(t)-1} \circ \mathrm{sign}(\mathbf{Y})$ is a matrix with the same size as $\mathbf{Y}$ in which the $(i,j)^{th}$ entry is $p(t) |Y_i(j)|^{p(t)-1} \mathrm{sign}(Y_i(j))$. The symbol $\circ$ refers to the Hadamard product (entrywise matrix product).

Thus, the gradient of $\hat{\psi}(\mathbf{W})$ is expressed as:

$$\nabla \hat{\psi}(\mathbf{W}(f)) = \mathbf{f}(\mathbf{Y}) \mathbf{X}^T \tag{11}$$

From (8) and (11), the natural gradient of $\hat{\psi}(\mathbf{W}_t)$ is:

$$\tilde{\nabla} \hat{\psi}(\mathbf{W}_t) = \mathbf{f}(\mathbf{Y}_t) \mathbf{Y}_t^T \mathbf{W}_t \tag{12}$$

The update equation of $\mathbf{W}_t$ for a frequency bin f is then:

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \mu \mathbf{G}_t \mathbf{W}_t \tag{13}$$

with $\mathbf{G}_t = \mathbf{f}(\mathbf{Y}_t)\mathbf{Y}_t^T$.

The convergence of the natural gradient is conditioned both by the initial coefficients $\mathbf{W}_0$ of the separation matrix and the step size of the update and it is quite difficult to choose the parameters that allow fast convergence without risking divergence. Douglas and Gupta [3] proposed to impose a scaling constraint to the separation matrix $\mathbf{W}_t$ to maintain a constant gradient magnitude for the algorithm. They assert that with this scaling and a fixed step size μ, the algorithm has fast convergence and excellent performance independently of the magnitude of $\mathbf{X}$ and $\mathbf{W}_0$. Applying this scaling constraint instead of the unit norm constraint, our update function becomes:

$$\mathbf{W}_{t+1} = c_t \mathbf{W}_t - \mu c_t^2 \mathbf{G}_t \mathbf{W}_t \tag{14}$$

with $c_t = \frac{1}{\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|g_t^{ij}\right|}$ and $g_t^{ij} = [\mathbf{G}_t]_{ij}$.

For the initialization of the separation matrix $\mathbf{W}_0$, we use a whitening process. The whitening is an important preprocessing in an overdetermined blind source separation algorithm as it allows to focus the energy of the received signals in the useful signal space. The separation matrix is initialized as follow:

$$\mathbf{W}_0 = \sqrt{\mathbf{D}_M^{-1}}\mathbf{E}_{:M}^H$$

where $\mathbf{D}_M$ is a matrix containing the first M rows and M columns of the matrix $\mathbf{D}$ and $\mathbf{E}_{:M}$ is the matrix containing the first M columns of the matrix $\mathbf{E}$. $\mathbf{D}$ and $\mathbf{E}$ are respectively the diagonal matrix and the unitary matrix of the singular value decomposition of the autocorrelation matrix of the received data $\mathbf{X}$.

The proposed algorithm is summarized as follow:

Algorithm 1 Parameterized l_p quasi-norm algorithm

1. *Input*: the output of the microphone array $\mathbf{x} = [\mathbf{x}(t_1), \ldots, \mathbf{x}(t_T)]$, the number of sources M and the optimization step μ
2. $\{\mathbf{X}(f,k)\}_{1\leq f\leq N_f, 1\leq k\leq N_T} = \mathrm{STFT}(\mathbf{x})$
3. for each frequency bin f,
   (a) initialize the separation matrix $\mathbf{W}_0(f)$ by a whitening process
   (b) $\mathbf{Y}_0(f,:) = \mathbf{W}_0(f)\mathbf{X}(f)$
   (c) for each iteration t,
      i. $\mathbf{f}(\mathbf{Y}_t(f,:)) = p(t)\left|\mathbf{Y}_t(f,:)\right|^{p(t)-1} \circ \mathrm{sign}(\mathbf{Y}_t(f,:))$
      ii. $\mathbf{G}_t(f) = \mathbf{f}(\mathbf{Y}_t(f,:))\mathbf{Y}_t^T(f,:)$
      iii. $c_t(f) = \frac{1}{\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|g_t^{ij}(f)\right|}$
      iv. $\mathbf{W}_{t+1}(f) = c_t(f)\mathbf{W}_t(f) - \mu c_t^2(f)\mathbf{G}_t(f)\mathbf{W}_t(f)$
      v. $\mathbf{Y}_{t+1}(f,:) = \mathbf{W}_{t+1}(f)\mathbf{X}(f)$
4. Permutation problem solving
5. *Output*: the estimated sources $\mathbf{y} = \mathrm{ISTFT}\left(\{\mathbf{Y}(f,k)\}_{1\leq f\leq N_f, 1\leq k\leq N_K}\right)$

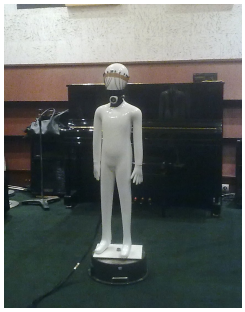

Figure 3: The dummy used to model the robot

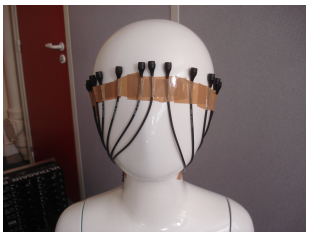

Figure 4: The 16 sensors microphone array

4. SIMULATIONS AND RESULTS

4.1 Experimental data

As we are in a robot audition context, we model the future robot by a child size dummy ($1m20$) for the sound acquisition process. We placed 16 sensors on the dummy head (*cf.* figure 3 and 4). The geometry of this microphone array is given by the figure 5. The output signals $\mathbf{x}(t)$ are the convolutions of 8 pairs of speech sources by the impulse responses $\{\mathbf{h}(l)\}_{0\leq l\leq L}$ measured from two angles of arrivals in a moderately reverberant room which reverberation time is $\mathrm{RT}_{30} = 300$ ms (*cf.* figure 6). In this experiment, all the 16 microphones are used and we are in a noiseless environment. More details of the blind source separation algorithms are given in table 1.

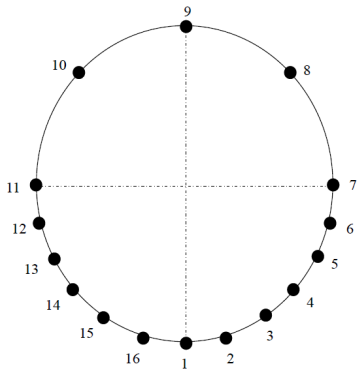

Figure 5: The geometry of the microphone array

Sampling frequency	16kHz
Analysis window	Hanning
Analysis window length	2048
Shift length	1024
μ	0.2
Direction of arrivals	0/90° and 0°/30°
Signals length	5s
Iteration number	500

Table 1: Implementation characteristics of the blind source separation algorithms

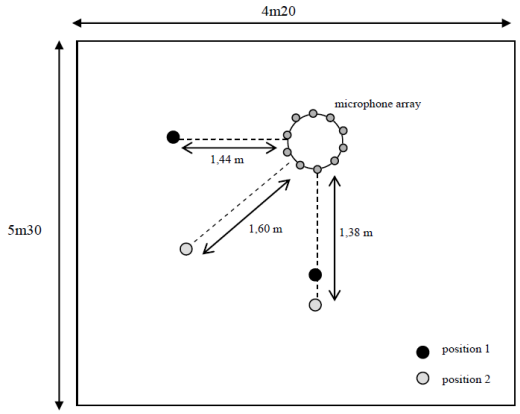

Figure 6: The position of the sound sources and the microphone array in the reverberant room

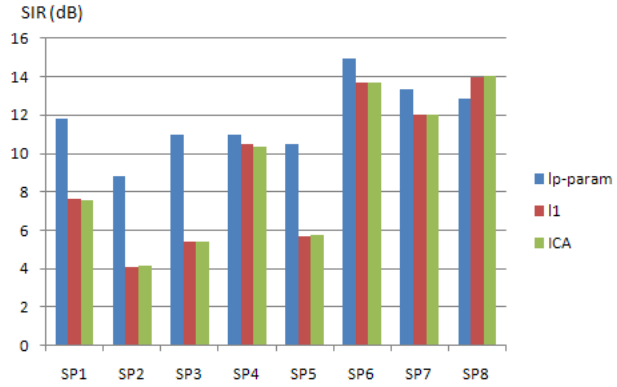

Figure 7: SIR of the blind source separation algorithms for the first position 0°/90° (SP*i* refers to the *ith* Source Pair)

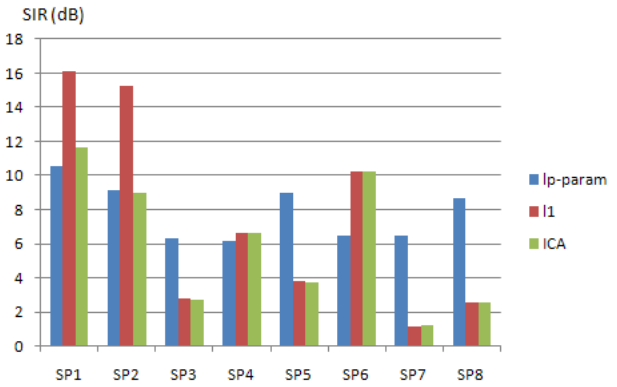

Figure 8: SIR of the blind source separation algorithms for the second position 0°/30° (SP*i* refers to the *ith* Source Pair)

4.2 Evaluation results

We evaluate the source separation algorithm by the Signal-to-Interference Ratio (SIR) using the *BSS-eval* toolbox [10]. The SIR is given by:

$$\text{SIR} = 10\log_{10}\frac{\left\|y_{target}\right\|^2}{\left\|e_{interf}\right\|^2} \tag{15}$$

where the estimated source signal y_j is decomposed as $y_j = y_{target} + e_{interf} + e_{noise} + e_{artif}$, y_{target} is a version of y_j modified by an allowed distortion and e_{interf} is the part of y_j perceived as coming from other unwanted sources $\{s_i\}_{i\neq j}$. We compare the parameterized l_p quasi-norm algorithm to the minimization of the l_1 norm and the scaled natural gradient based ICA algorithm in [3]. The minimization of the l_1 norm is basically the same algorithm than the one presented in section 3 with a fixed parameter $p = 1$.

Figures 7 and 8 show the SIR results for the 3 algorithms in the first and second positions of the sources and for 8 source pairs. The parameterized l_p quasi-norm algorithm, named lp-param in the results figures, has good results and is comparable to the ICA algorithm and better in some cases.

Figure 9 shows the sparsity curves of the estimated sources through the iterations measured by the Gini index which is a good sparsity measure for speech [7]. The Gini index is bounded by 0 and 1: the closer the index is to 1, the sparser is the signal. We notice that the sparsity of the estimated sources grows through the iterations and converges quickly to its optimal value (in less than 50 iterations in this context). As the source pairs used for the separation are different, we should not expect a correlation between the SIR results and the value of the Gini index for a pair of sources. For different pairs of sources, the value of the Gini index of the estimated sources is closely related to the Gini index of the original sources and not to the separation performance.

One question to ask is how evolves the algorithm with a fixed parameter $p < 1$ comparing to a parameterized one. Figures 10 and 11 show some results in the case of a fixed $p < 1$. For each value of p, we run the l_p norm algorithm with the specified number of iteration. Figure 10 is the SIR of a pair of source in the first position and it shows that the best SIR is obtained for $p = 0.3$, better than the SIR with the parameterized l_p quasi-norm algorithm. The same observation can be made for the figure 11, which represents the SIR for a pair of sources in the second position for $p < 1$. In this case, the best SIR is obtained in this case is for $p = 0.9$.

If we bypass the stability issues that can occur for a fixed $p < 1$, we can obtain better results than the parameterized l_p quasi-norm algorithm, l_1 and ICA for an optimal value of p. The main problem in this case is the choice of this optimal value of p. The parameterized l_p quasi-norm is then a good solution if we want to harder the sparsity constraint without making an heuristic choice of the parameter p.

5. CONCLUSION

We propose a new blind source separation algorithm based on the minimization of a parameterized l_p norm (a quasi-norm if $0 < p < 1$). We introduced a decrease of the parameter p according to a sigmoid function which makes the sparsity constraint harder through the iterations. The decrease according to a sigmoid allows us to avoid the divergence of the algorithm when changing the l_p norm from an iteration to another. The results show that the proposed algorithm has good performance comparing to the minimization of the l_1

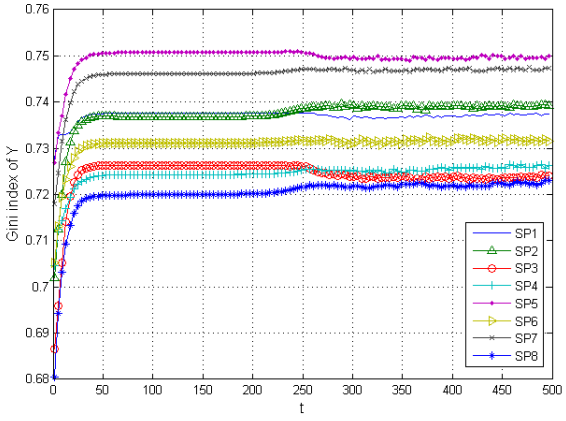

Figure 9: The Gini index of the estimated sources through the iterations in the first recording situation

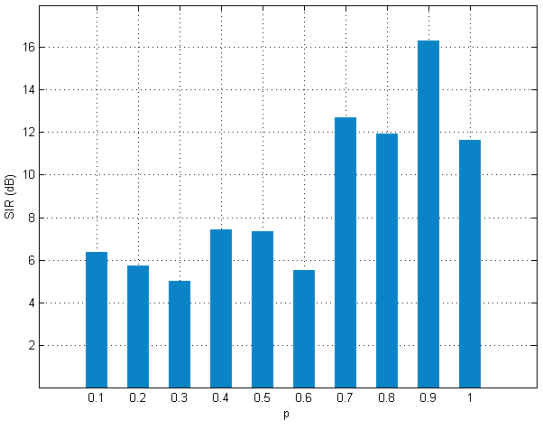

Figure 11: SIR for SP6 in the first position (0°/30°) with p varying from 0.1 to 1 (same p through the iterations)

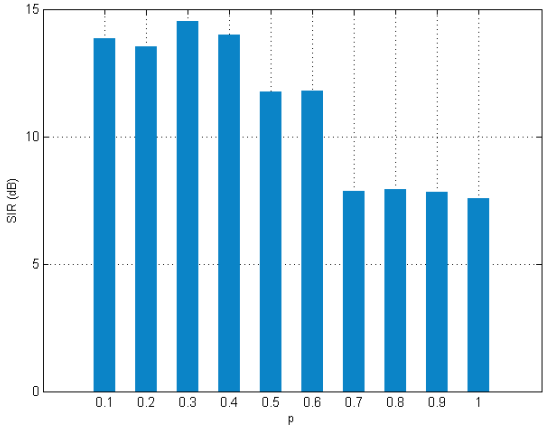

Figure 10: SIR for SP1 in the first position (0°/90°) with p varying from 0.1 to 1 (same p through the iterations)

norm and ICA.

REFERENCES

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

[2] Shun-Ichi Amari. Natural gradient works efficiently in learning. In *Neural Computation*, volume 10, pages 251–276, Cambridge, MA, USA, 1998. MIT Press.

[3] S.C. 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*, volume 2, pages II–637 –II–640, Apr. 2007.

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

[5] H. Nakajima, K. Nakadai, Y. Hasegawa, and H. Tsujino. High performance sound source separation adaptable to environmental changes for robot audition. *IEEE/RSJ International Conference on Intelligent Robots and Systems, 2008, IROS 2008, Nice*, pages 2165–2171, Sept. 2008.

[6] Romeo project. www.projetromeo.com.

[7] Scott Rickard and Maurice Fallon. The gini index of speech. *Conference on Information Sciences and Systems, Princeton*, March 2004.

[8] H. Saruwatari, Y. Mori, T. Takatani, S. Ukai, K. Shikano, T. Hiekata, and T. Morita. Two-stage blind source separation based on ica and binary masking for real-time robot audition system. *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2005, Edmonton, Alberta*, pages 2303–2308, 2005.

[9] Y. Tamai, Y. Sasaki, S. Kagami, and H. Mizoguchi. Three ring microphone array for 3d sound localization and separation for mobile robot audition. *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2005, Edmonton, Alberta*, pages 4172–4177, Aug. 2005.

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

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

[12] S. Yamamoto, K. Nakadai, M. Nakano, H. Tsujino, J.-M. Valin, K. Komatani, T. Ogata, and H.G. Okuno. Design and implementation of a robot audition system for automatic speech recognition of simultaneous speech. *IEEE Workshop on Automatic Speech Recognition Understanding, ASRU 2007, Kyoto*, pages 111–116, 2007.