

# Enhanced Robot Audition Based on Microphone Array Source Separation with Post-Filter

Jean-Marc Valin, Jean Rouat, François Michaud

LABORIUS, Department of Electrical Engineering and Computer Engineering

Université de Sherbrooke, Sherbrooke (Quebec) CANADA, J1K 2R1

{Jean-Marc.Valin, Jean.Rouat, Francois.Michaud}@USherbrooke.ca

**Abstract**—We propose a system that gives a mobile robot the ability to separate simultaneous sound sources. A microphone array is used along with a real-time dedicated implementation of Geometric Source Separation and a post-filter that gives us a further reduction of interferences from other sources. We present results and comparisons for separation of multiple non-stationary speech sources combined with noise sources. The main advantage of our approach for mobile robots resides in the fact that both the frequency-domain Geometric Source Separation algorithm and the post-filter are able to adapt rapidly to new sources and non-stationarity. Separation results are presented for three simultaneous interfering speakers in the presence of noise. A reduction of log spectral distortion (LSD) and increase of signal-to-noise ratio (SNR) of approximately 10 dB and 14 dB are observed.

## I. INTRODUCTION

Our hearing sense allows us to perceive all kinds of sounds (speech, music, phone ring, closing a door, etc.) in our world, whether we are moving or not. To operate in human and natural settings, autonomous mobile robots should be able to do the same. This requires the robots not just to detect sounds, but also to localise their origin, separate the different sound sources (since sounds may occur simultaneously), and process all of this data to extract useful information about the world.

Even though artificial hearing would be an important sensing capability for autonomous systems, the research topic is still in its infancy. Only a few robots are using hearing capabilities: SAIL [1] uses one microphone to develop online audio-driven behaviors; ROBITA [2] uses two microphones to follow a conversation between two persons; SIG [3], [4], [5] uses one pair of microphones to collect sound from the external world, and another pair placed inside the head to collect internal sounds (caused by motors) for noise cancellation; Sony SDR-4X has seven microphones; a service robot uses eight microphones organised in a circular array to do speech enhancement and recognition [6]. Even though robots are not limited to only two ears, they still have not shown the capabilities of the human hearing sense.

We address the problem of isolating sound sources from the environment. The human hearing sense is very good at focusing on a single source of interest despite all kinds of

interferences. We generally refer to this ability as the *cocktail party effect*, where a human listener is able to follow a conversation even when several people are speaking at the same time. For a mobile robot, it would mean being able to separate all sound sources present in the environment at any moment.

Working toward that goal, our interest in this paper is to describe a two-step approach for performing sound source separation on a mobile robot equipped with an array of eight low-cost microphones. The initial step consists of a linear separation based on a simplified version of the Geometric Source Separation approach proposed by Parra and Alvino [7] with a faster stochastic gradient estimation and shorter time frames estimations. The second step is a generalisation of beamformer post-filtering [8], [9] for multiple sources and uses adaptive spectral estimation of background noise and interfering sources to enhance the signal produced during the initial separation. The novelty of this post-filter resides in the fact that, for each source of interest, the noise estimate is decomposed into stationary and transient components assumed to be due to leakage between the output channels of the initial separation stage.

The paper is organised as follows. Section II gives an overview of the system. Section III presents the linear separation algorithm and Section IV describes the proposed post-filter. Results are presented in Section V, followed by the conclusion.

## II. SYSTEM OVERVIEW

The proposed sound separation algorithm as shown in Figure 1 is composed of three parts:

- 1) A microphone array;
- 2) A linear source separation algorithm (LSS) implemented as a variant of the Geometric Source Separation (GSS) algorithm;
- 3) A multi-channel post-filter.

The microphone array is composed of a number of omnidirectional elements mounted on the robot. The microphone signals are combined linearly in a first-pass separation algorithm. The output of this initial separation is then enhanced by a (non-linear) post-filter designed to optimally attenuate the remaining noise and interference from other sources.

We assume that these sources are detected and localised by an algorithm such as [10] (our approach is not specific to

<sup>0</sup>©2004 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

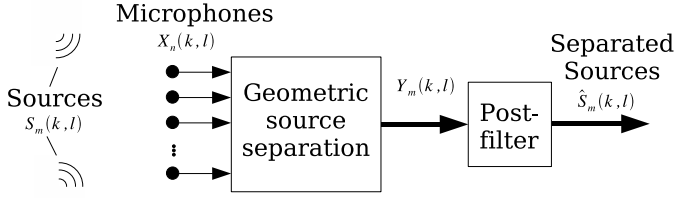


Figure 1. Overview of the separation system

any localisation algorithm). We also assume that sources may appear, disappear or move at any time. It is thus necessary to maximise the adaptation rate for both the LSS and the multi-channel post-filter. Mobile robotics also imposes real-time constraints: the algorithmic delay must be kept small and the complexity must be low enough for the data to be processed in real-time on a conventional processor.

### III. LINEAR SOURCE SEPARATION

The LSS algorithm we propose in this section is based on the Geometric Source Separation (GSS) approach proposed by Parra and Alvino [7]. Unlike the Linearly Constrained Minimum Variance (LCMV) beamformer that minimises the output power subject to a distortionless constraint, GSS explicitly minimises cross-talk, leading to faster adaptation. The method is also interesting for use in the mobile robotics context because it allows easy addition and removal of sources. Using some approximations described in Subsection III-B, it is also possible to implement separation with relatively low complexity (i.e. complexity that grows linearly with the number of microphones).

#### A. Geometric Source Separation

The method operates in the frequency domain. Let  $S_m(k, \ell)$  be the real (unknown) sound source  $m$  at time frame  $\ell$  and for discrete frequency  $k$ . We denote as  $\mathbf{s}(k, \ell)$  the vector corresponding to the sources  $S_m(k, \ell)$  and matrix  $\mathbf{A}(k)$  is the transfer function leading from the sources to the microphones. The signal received at the microphones is thus given by:

$$\mathbf{x}(k, \ell) = \mathbf{A}(k)\mathbf{s}(k, \ell) + \mathbf{n}(k, \ell) \quad (1)$$

where  $\mathbf{n}(k, \ell)$  is the non-coherent background noise received at the microphones. The matrix  $\mathbf{A}(k)$  can be estimated using the result of a sound localisation algorithm. Assuming that all transfer functions have unity gain, the elements of  $\mathbf{A}(k)$  can be expressed as:

$$a_{ij}(k) = e^{-j2\pi k \delta_{ij}} \quad (2)$$

where  $\delta_{ij}$  is the time delay (in samples) to reach microphone  $i$  from source  $j$ .

The separation result is then defined as  $\mathbf{y}(k, \ell) = \mathbf{W}(k, \ell)\mathbf{x}(k, \ell)$ , where  $\mathbf{W}(k, \ell)$  is the separation matrix that must be estimated. This is done by providing two constraints (the index  $\ell$  is omitted for the sake of clarity):

- 1) Decorrelation of the separation algorithm outputs, expressed as  $\mathbf{R}_{\mathbf{yy}}(k) - \text{diag}[\mathbf{R}_{\mathbf{yy}}(k)] = \mathbf{0}^1$ .
- 2) The geometric constraint  $\mathbf{W}(k)\mathbf{A}(k) = \mathbf{I}$ , which ensures unity gain in the direction of the source of interest and places zeros in the direction of interferences.

In theory, constraint 2) could be used alone for separation (the method is referred to as LS-C2 in [7]), but in practice, the method does not take into account reverberation or errors in localisation. It is also subject to instability if  $\mathbf{A}(k)$  is not invertible at a specific frequency. When used together, constraints 1) and 2) are too strong. For this reason, we propose “soft” constraints that are a combination of 1) and 2) in the context of a gradient descent algorithm.

Two cost functions are created by computing the square of the error associated with constraints 1) and 2). These cost functions are respectively defined as:

$$J_1(\mathbf{W}(k)) = \|\mathbf{R}_{\mathbf{yy}}(k) - \text{diag}[\mathbf{R}_{\mathbf{yy}}(k)]\|^2 \quad (3)$$

$$J_2(\mathbf{W}(k)) = \|\mathbf{W}(k)\mathbf{A}(k) - \mathbf{I}\|^2 \quad (4)$$

where the matrix norm is defined as  $\|\mathbf{M}\|^2 = \text{trace}[\mathbf{M}\mathbf{M}^H]$  and is equal to the sum of the square of all elements in the matrix. The gradient of the cost functions with respect to  $\mathbf{W}(k)$  is equal to [7]:

$$\frac{\partial J_1(\mathbf{W}(k))}{\partial \mathbf{W}^*(k)} = 4\mathbf{E}(k)\mathbf{W}(k)\mathbf{R}_{\mathbf{xx}}(k) \quad (5)$$

$$\frac{\partial J_2(\mathbf{W}(k))}{\partial \mathbf{W}^*(k)} = 2[\mathbf{W}(k)\mathbf{A}(k) - \mathbf{I}]\mathbf{A}(k) \quad (6)$$

where  $\mathbf{E}(k) = \mathbf{R}_{\mathbf{yy}}(k) - \text{diag}[\mathbf{R}_{\mathbf{yy}}(k)]$ .

The separation matrix  $\mathbf{W}(k)$  is then updated as follows:

$$\mathbf{W}^{n+1}(k) = \mathbf{W}^n(k) - \mu \left[ \alpha(k) \frac{\partial J_1(\mathbf{W}(k))}{\partial \mathbf{W}^*(k)} + \frac{\partial J_2(\mathbf{W}(k))}{\partial \mathbf{W}^*(k)} \right] \quad (7)$$

where  $\alpha(f)$  is an energy normalisation factor equal to  $\|\mathbf{R}_{\mathbf{xx}}(k)\|^{-2}$  and  $\mu$  is the adaptation rate.

#### B. Stochastic Gradient Adaptation

The difference between our algorithm and the original GSS algorithm described in [7] is that instead of estimating the correlation matrices  $\mathbf{R}_{\mathbf{xx}}(k)$  and  $\mathbf{R}_{\mathbf{yy}}(k)$  on several seconds of data, our approach uses instantaneous estimations. This is analogous to the approximation made in the Least Mean Square (LMS) adaptive filter [11]. We thus assume that:

$$\mathbf{R}_{\mathbf{xx}}(k) = \mathbf{x}(k)\mathbf{x}(k)^H \quad (8)$$

$$\mathbf{R}_{\mathbf{yy}}(k) = \mathbf{y}(k)\mathbf{y}(k)^H \quad (9)$$

It is then possible to rewrite the gradient  $\frac{\partial J_1(\mathbf{W}(k))}{\partial \mathbf{W}^*(k)}$  as:

$$\frac{\partial J_1(\mathbf{W}(k))}{\partial \mathbf{W}^*(k)} = 4[\mathbf{E}(k)\mathbf{W}(k)\mathbf{x}(k)]\mathbf{x}(k)^H \quad (10)$$

which only requires matrix-by-vector products, greatly reducing the complexity of the algorithm. The normalisation factor

<sup>1</sup>Assuming non-stationary sources, second order statistics are sufficient for ensuring independence of the separated sources.

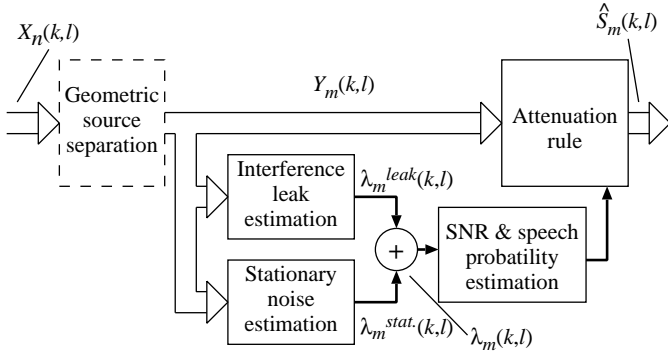


Figure 2. Overview of the post-filter.  
 $X_n(k, \ell)$ ,  $n = 0 \dots N-1$ : Microphone inputs,  $Y_m(k, \ell)$ ,  $m = 0 \dots M-1$ :  
Inputs to the post-filter,  $\hat{S}_m(k, \ell) = G_m(k, \ell)Y_m(k, \ell)$ ,  $m = 0 \dots M-1$ :  
Post-filter outputs.

$\alpha(k)$  can also be simplified as  $\left[\|\mathbf{x}(k)\|^2\right]^{-2}$ . From this work, the instantaneous estimation of the correlation has not shown any reduction in accuracy and furthermore eases real-time integration.

### C. Initialisation

The fact that sources can appear or disappear at any time imposes constraints on the initialisation of the separation matrix  $\mathbf{W}(k)$ . The initialisation must provide the following:

- The initial weights for a new source;
- Acceptable separation (before adaptation).

Furthermore, when a source appears or disappears, other sources must be unaffected.

One easy way to satisfy both constraints is to initialise the column of  $\mathbf{W}(k)$  corresponding to the new source  $m$  as:

$$w_{m,i}(k) = \frac{a_{i,m}(k)}{N} \quad (11)$$

This initialisation is equivalent to a delay-and-sum beamformer, and is referred to as the I1 initialisation method in [7].

## IV. MULTI-CHANNEL POST-FILTER

In order to enhance the output of the GSS algorithm presented in Section III, we derive a frequency-domain post-filter that is based on the optimal estimator originally proposed by Ephraim and Malah [12], [13]. Several approaches to microphone array post-filtering have been proposed in the past. Most of these post-filters address reduction of stationary background noise [14], [15]. Recently, a multi-channel post-filter taking into account non-stationary interferences was proposed by Cohen [8]. The novelty of our approach resides in the fact that, for a given channel output of the GSS, the transient components of the corrupting sources is assumed to be due to leakage from the other channels during the GSS process. Furthermore, for a given channel, the stationary and the transient components are combined into a single noise estimator used for noise suppression, as shown in Figure 2.

For this post-filter, we consider that all interferences (except the background noise) are localised (detected by the localisation algorithm) sources and we assume that the leakage between channels is constant. This leakage is due to reverberation, localisation error, differences in microphone frequency responses, near-field effects, etc.

Section IV-A describes the estimation of noise variances that are used to compute the weighting function  $G_m$  by which the outputs  $Y_m$  of the LSS is multiplied to generate a cleaned signal whose spectrum is denoted  $\hat{S}_m$ .

### A. Noise estimation

The noise variance estimation  $\lambda_m(k, \ell)$  is expressed as:

$$\lambda_m(k, \ell) = \lambda_m^{stat.}(k, \ell) + \lambda_m^{leak}(k, \ell) \quad (12)$$

where  $\lambda_m^{stat.}(k, \ell)$  is the estimate of the stationary component of the noise for source  $m$  at frame  $\ell$  for frequency  $k$ , and  $\lambda_m^{leak}(k, \ell)$  is the estimate of source leakage.

We compute the stationary noise estimate  $\lambda_m^{stat.}(k, \ell)$  using the Minima Controlled Recursive Average (MCRA) technique proposed by Cohen [16].

To estimate  $\lambda_m^{leak}$  we assume that the interference from other sources is reduced by a factor  $\eta$  (typically  $-10 \text{ dB} \leq \eta \leq -5 \text{ dB}$ ) by the separation algorithm (LSS). The leakage estimate is thus expressed as:

$$\lambda_m^{leak}(k, \ell) = \eta \sum_{i=0, i \neq m}^{M-1} Z_i(k, \ell) \quad (13)$$

where  $Z_m(k, \ell)$  is the smoothed spectrum of the  $m^{th}$  source,  $Y_m(k, \ell)$ , and is recursively defined (with  $\alpha_s = 0.7$ ) as:

$$Z_m(k, \ell) = \alpha_s Z_m(k, \ell - 1) + (1 - \alpha_s) Y_m(k, \ell) \quad (14)$$

### B. Suppression rule in the presence of speech

We now derive the suppression rule under  $H_1$ , the hypothesis that speech is present. From here on, unless otherwise stated, the  $m$  index and the  $\ell$  arguments are omitted for clarity and the equations are given for each  $m$  and for each  $\ell$ .

The proposed noise suppression rule is based on minimum mean-square error (MMSE) estimation of the spectral amplitude in the loudness domain,  $|X(k)|^{1/2}$ . The choice of the loudness domain over the spectral amplitude [12] or log-spectral amplitude [13] is motivated by better results obtained using this technique, mostly when dealing with speech presence uncertainty (Section IV-C).

The loudness-domain amplitude estimator is defined by:

$$\hat{A}(k) = (E[|S(k)|^\alpha |Y(k)|])^{\frac{1}{\alpha}} = G_{H_1}(k) |Y(k)| \quad (15)$$

where  $\alpha = 1/2$  for the loudness domain and  $G_{H_1}(k)$  is the spectral gain assuming that speech is present.

The spectral gain for arbitrary  $\alpha$  is derived from Equation 13 in [13]:

$$G_{H_1}(k) = \frac{\sqrt{v(k)}}{\gamma(k)} \left[ \Gamma\left(1 + \frac{\alpha}{2}\right) M\left(-\frac{\alpha}{2}; 1; -v(k)\right) \right]^{\frac{1}{\alpha}} \quad (16)$$

where  $M(a; c; x)$  is the confluent hypergeometric function,  $\gamma(k) \triangleq |Y(k)|^2 / \lambda(k)$  and  $\xi(k) \triangleq E[|S(k)|^2] / \lambda(k)$  are respectively the *a posteriori* SNR and the *a priori* SNR. We also have  $v(k) \triangleq \gamma(k)\xi(k) / (\xi(k) + 1)$  [12].

The *a priori* SNR  $\xi(k)$  is estimated recursively as:

$$\begin{aligned} \hat{\xi}(k, \ell) &= \alpha_p G_{H_1}^2(k, \ell - 1) \gamma(k, \ell - 1) \\ &+ (1 - \alpha_p) \max\{\gamma(k, \ell) - 1, 0\} \end{aligned} \quad (17)$$

using the modifications proposed in [16] to take into account speech presence uncertainty.

### C. Optimal gain modification under speech presence uncertainty

In order to take into account the probability of speech presence, we derive the estimator for the loudness domain:

$$\hat{A}(k) = (E[A^\alpha(k)|Y(k)])^{\frac{1}{\alpha}} \quad (18)$$

Considering  $H_1$ , the hypothesis of speech presence for source  $m$ , and  $H_0$ , the hypothesis of speech absence, we obtain:

$$\begin{aligned} E[A^\alpha(k)|Y(k)] &= p(k)E[A^\alpha(k)|H_1, Y(k)] \\ &+ [1 - p(k)]E[A^\alpha(k)|H_0, Y(k)] \end{aligned} \quad (19)$$

where  $p(k)$  is the probability of speech at frequency  $k$ .

The optimally modified gain is thus given by:

$$G(k) = [p(k)G_{H_1}^\alpha(k) + (1 - p(k))G_{min}^\alpha]^\frac{1}{\alpha} \quad (20)$$

where  $G_{H_1}(k)$  is defined in (16), and  $G_{min}$  is the minimum gain allowed when speech is absent. Unlike the log-amplitude case, it is possible to set  $G_{min} = 0$  without running into problems. For  $\alpha = 1/2$ , this leads to:

$$G(k) = p^2(k)G_{H_1}(k) \quad (21)$$

Setting  $G_{min} = 0$  means that there is no arbitrary limit on attenuation. Therefore, when the signal is certain to be non-speech, the gain can tend toward zero. This is especially important when the interference is also speech since, unlike stationary noise, residual babble noise always results in musical noise.

The probability of speech presence is computed as:

$$p(k) = \left\{ 1 + \frac{\hat{q}(k)}{1 - \hat{q}(k)} (1 + \xi(k)) \exp(-v(k)) \right\}^{-1} \quad (22)$$

where  $\hat{q}(k)$  is the *a priori* probability of speech presence for frequency  $k$  and is defined as:

$$\hat{q}(k) = 1 - P_{local}(k)P_{global}(k)P_{frame} \quad (23)$$

where  $P_{local}(k)$ ,  $P_{global}(k)$  and  $P_{frame}$  are defined in [16] and correspond respectively to a speech measurement on the current frame for a local frequency window, a larger frequency and for the whole frame.

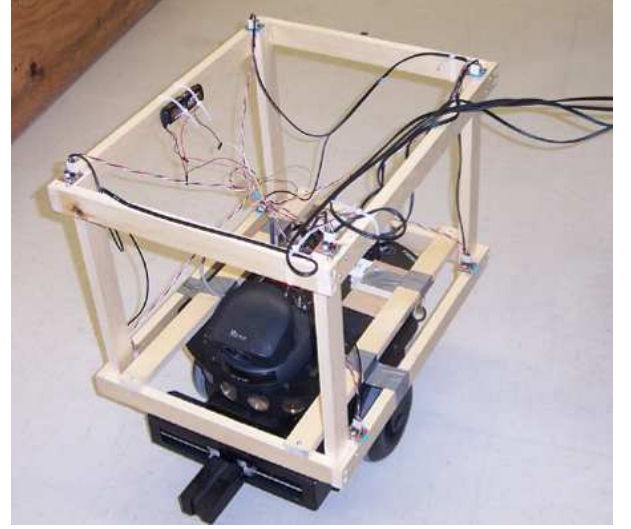


Figure 3. Pioneer 2 robot with an array of eight microphones

### D. Initialisation

When a new source appears, post-filter state variables need to be initialised. Most of these variables may safely be set to zero. The exception is  $\lambda_m^{stat.}(k, \ell_0)$ , the initial stationary noise estimation for source  $m$ . The MCRA algorithm requires several seconds to produce its first estimate for source  $m$ , so it is necessary to find another way to estimate the background noise until a better estimate is available. This initial estimate is thus computed using noise estimations at the microphones. Assuming the delay-and-sum initialisation of the weights from Equation 11, the initial background noise estimate is thus:

$$\lambda_m^{stat.}(k, \ell_0) = \frac{1}{N^2} \sum_{n=0}^{N-1} \sigma_{x_n}^2(k) \quad (24)$$

where  $\sigma_{x_n}^2(k)$  is the noise estimation for microphone  $n$ .

## V. RESULTS

Our system is evaluated on a Pioneer 2 robot, on which an array of eight microphones is installed. In order to test the system, three voices (two female, one male) were recorded separately, in a quiet environment. The background noise was recorded on the robot and includes the room ventilation and the internal robot fans. All four signals were recorded using the same microphone array and subsequently mixed together. This procedure was required in order to compute the distance measures (such as SNR) presented in this section. It is worth noting that although the signals were mixed artificially, the result still represents real conditions with background noise, interfering sources, and reverberation.

In evaluating our source separation system, we use the conventional signal-to-noise ratio (SNR) and the log spectral

Table I  
SIGNAL-TO-NOISE RATIO (SNR) FOR EACH OF THE THREE SEPARATED SOURCES.

SNR (dB)	female 1	female 2	male 1
Microphone inputs	-1.8	-3.7	-5.2
Delay-and-sum	7.3	4.4	-1.2
GSS	9.0	6.0	3.7
GSS+single channel	9.9	6.9	4.5
GSS+multi-channel	12.1	9.5	9.4

Table II  
LOG-SPECTRAL DISTORTION (LSD) FOR EACH OF THE THREE SEPARATED SOURCES.

LSD (dB)	female 1	female 2	male 1
Microphone inputs	17.5	15.9	14.8
Delay-and-sum	15.8	15.0	15.1
GSS	15.0	14.2	14.2
GSS+single channel	9.7	9.5	10.4
GSS+multi-channel	6.5	6.8	7.4

distortion (LSD), that is defined as:

$$LSD = \frac{1}{L} \sum_{\ell=0}^{L-1} \left[ \frac{1}{K} \sum_{k=0}^{K-1} \left( 10 \log_{10} \frac{|S(k, \ell)|^2 + \epsilon}{|\hat{S}(k, \ell)|^2 + \epsilon} \right)^2 \right]^{\frac{1}{2}} \quad (25)$$

where  $L$  is the number of frames,  $K$  is the number of frequency bins and  $\epsilon$  is meant to prevent extreme values for spectral regions of low energy.

Tables I and II compare the results obtained for different configurations: unprocessed microphone inputs, delay-and-sum algorithm, GSS algorithm, GSS algorithm with single-channel post-filter, and GSS algorithm with multi-channel post-filter (proposed). It is worth noting that the delay-and-sum algorithm corresponds to the initial value of the separation matrix provided to our algorithm. While it is clear that GSS performs better than delay-and-sum, the latter still provides acceptable separation capabilities. These results also show that our multi-channel post-filter provides a significant improvement over both the single-channel post-filter and plain GSS.

The signals amplitude for the first source (female) are shown in Figure 5 and the spectrograms are shown in Figure 4. Even though the task involves non-stationary interference with the same frequency content as the signal of interest, we observe that our post-filter (unlike the single-channel post-filter) is able to remove most of the interference, while not causing excessive distortion to the signal of interest. Informal subjective evaluation has confirmed that the post-filter has a positive impact on both quality and intelligibility of the speech<sup>2</sup>.

## VI. CONCLUSION

In this paper we describe a microphone array linear source separator and a post-filter in the context of multiple and

<sup>2</sup>Audio signals and spectrograms for all three sources are available at: <http://www.speech.org/~jm/phd/separation/>

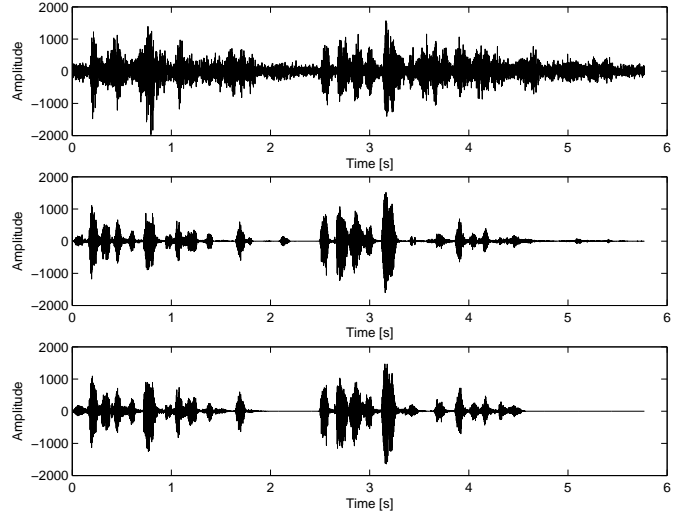


Figure 5. Signal amplitude for separation of first source (female voice). top: signal at one microphone. middle: system output. bottom: reference (clean) signal.

simultaneous sound sources. The linear source separator is based on a simplification of the geometric source separation algorithm that performs instantaneous estimation of the correlation matrix  $\mathbf{R}_{xx}(k)$ . The post-filter is based on a loudness-domain MMSE estimator in the frequency domain with a noise estimate that is computed as the sum of a stationary noise estimate and an estimation of leakage from the geometric source separation algorithm. The proposed post-filter is also sufficiently general to be used in addition to most linear source separation algorithms.

Experimental results show a reduction in log spectral distortion of up to 11 dB and an increase of the signal-to-noise ratio of 14dB compared to the noisy signal inputs. Preliminary perceptive test and visual inspection of spectrograms show us that the distortions introduced by the system are acceptable to most listeners.

A possible next step for this work would consist of directly optimizing the separation results for speech recognition accuracy. Also, a possible improvement to the algorithm would be to derive a method that automatically adapts the leakage coefficient  $\eta$  to track the leakage of the GSS algorithm.

## ACKNOWLEDGMENT

François Michaud holds the Canada Research Chair (CRC) in Mobile Robotics and Autonomous Intelligent Systems. This research is supported financially by the CRC Program, the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canadian Foundation for Innovation (CFI). Special thanks to Dominic Létourneau and Serge Caron for their help in this work.

## REFERENCES

- [1] Y. Zhang and J. Weng, "Grounded auditory development by a developmental robot," in *Proc. INNS/IEEE International Joint Conference of Neural Networks*, 2001, pp. 1059–1064.

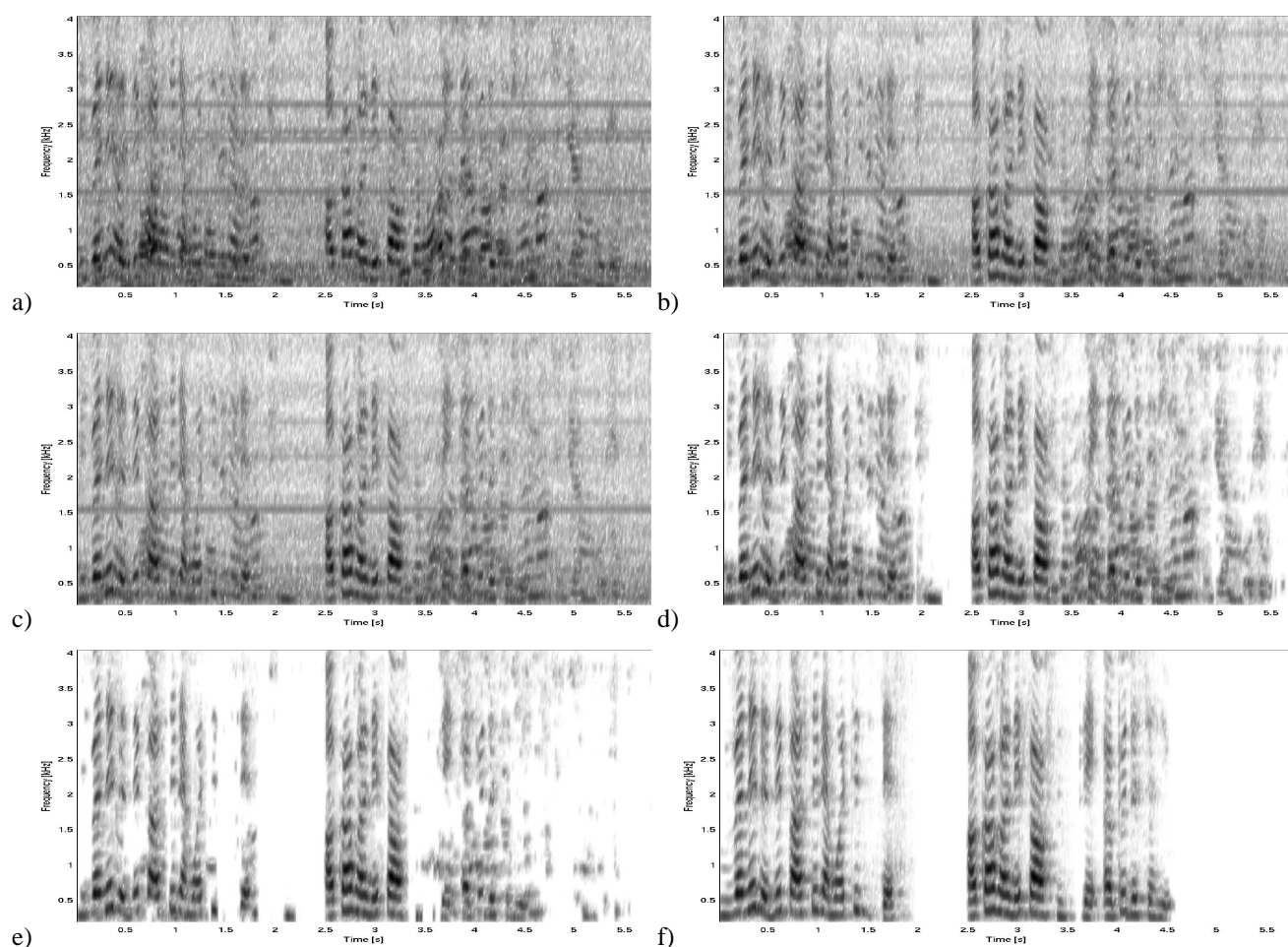


Figure 4. Spectrograms for separation of first source (female voice): a) signal at one microphone, b) delay-and-sum beamformer, c) GSS output, d) GSS with single-channel post-filter, e) GSS with multi-channel post-filter, f) reference (clean) signal.

- [2] Y. Matsusaka, T. Tojo, S. Kubota, K. Furukawa, D. Tamiya, K. Hayata, Y. Nakano, and T. Kobayashi, "Multi-person conversation via multi-modal interface - a robot who communicate with multi-user," in *Proc. EUROSPEECH*, 1999, pp. 1723–1726.
- [3] K. Nakadai, H. G. Okuno, and H. Kitano, "Real-time sound source localization and separation for robot audition," in *Proc. IEEE International Conference on Spoken Language Processing*, 2002, pp. 193–196.
- [4] H. G. Okuno, K. Nakadai, and H. Kitano, "Social interaction of humanoid robot based on audio-visual tracking," in *Proc. of Eighteenth International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, 2002, pp. 725–735.
- [5] K. Nakadai, D. Matsuura, H. G. Okuno, and H. Kitano, "Applying scattering theory to robot audition system: Robust sound source localization and extraction," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2003, pp. 1147–1152.
- [6] C. Choi, D. Kong, J. Kim, and S. Bang, "Speech enhancement and recognition using circular microphone array for service robots," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2003, pp. 3516–3521.
- [7] L. C. Parra and C. V. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 6, pp. 352–362, 2002.
- [8] I. Cohen and B. Berdugo, "Microphone array post-filtering for non-stationary noise suppression," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002, pp. 901–904.
- [9] J.-M. Valin, J. Rouat, and F. Michaud, "Microphone array post-filter for separation of simultaneous non-stationary sources," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004.
- [10] J.-M. Valin, F. Michaud, B. Hadjou, and J. Rouat, "Localization of simultaneous moving sound sources for mobile robot using a frequency-domain steered beamformer approach," in *Proc. IEEE International Conference on Robotics and Automation*, 2004.
- [11] S. Haykin, *Adaptive Filter Theory*, 4th ed. Prentice Hall, 2002.
- [12] Y. Ephraim and D. Malah, "Speech enhancement using minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, 1984.
- [13] —, "Speech enhancement using minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-33, no. 2, pp. 443–445, 1985.
- [14] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1988, pp. 2578–2581.
- [15] I. McCowan and H. Bourlard, "Microphone array post-filter for diffuse noise field," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 2002, pp. 905–908.
- [16] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 2, pp. 2403–2418, 2001.