Music Source Separation via Sparsified Dictionaries vs. Parametric Models

Mahdi Triki[†], Dirk T.M. Slock[*]
[†] CNRS, Communication Systems Laboratory
[*]Eurecom Institute
2229 route des Crêtes, B.P. 193, 06904 Sophia Antipolis Cedex, FRANCE
Email: {triki,slock}@eurecom.fr

***Abstract*— In the framework of audio signal analysis, there have been recent significant advances in two directions: sparse and structured representation. In fact, sparse decompositions of audio signals are shown to be effective, and appear to be extremely useful in many signal processing applications: compression, source separation, noise reduction... A second point of view tries to take advantage of the harmonic structure of the audio signal. It models a note signal as a periodic signal with (slow) global variation of amplitude (reflecting attack, sustain, decay) and frequency (limited time warping).**

In this paper, we compare the two approaches through experiments involving various audio signals. We consider particularly application of the two approaches to noise reduction, and underdetermined Blind Source Separation.

I. Introduction

The majority of blind separation algorithms are based on the theory of Independent Component Analysis. The idea is to estimate the inverse mixing matrix using statistical independence of source signals. However, one area of research in Blind Source Separation (BSS), the Underdetermined BSS, is relatively unexplored. It refers to the case when there are less mixtures than sources. The underdetermined BSS poses a challenge because the mixing matrix is not invertible and the traditional ICA methods do not work. And, contrary to most blind separation algorithms, the source extraction itself requires additional assumptions on the source statistics or structure. Several approaches in the literature are proposed to solve the problem exploiting essentially the time-frequency sparcity of the source signals [1], [2]. So that, they decompose the degenerate blind separation problem into several overdetermined problems.

On the other hand, for the representation of audio signals, there have been recent significant advances in two directions: sparse and structured representations. In fact, audio signals contain superimposed structures such as transients and stationary parts, or multiple notes and instruments; and have been shown to have sparse decompositions in a variety of time-frequency dictionaries. The goal is to decompose the audio signal onto a small number of basis functions, called "atoms" (typically time-frequency atoms, such as local cosines, or time-scale atoms, such as wavelets). The fundamental problem is that the bigger the dictionary (i.e. the more redundant), the more likely to have a good match between the signal and the atoms; but the larger the set of possible solutions. As computing the optimal solution is an NP-hard problem, sub-optimal greedy strategies are introduced to decompose, in an iterative fashion, any signal into a linear expansion of waveforms belonging to the given dictionary [3], [5], [6], [7]. in the second point of view, one tries to take advantage of the harmonically structure of the audio signal. For treating harmonic structured signals, Parks et al. consider the estimation of pure periodic signals with period equal to an integer number of samples [8], [9]. In these references, the authors propose a Maximum Likelihood approach to analyze pure periodic signals. They show that the resulting procedure can be interpreted as a signal projection onto suitable subspaces. In [10], [11], we extended the results of those references, and we merged the modulated sinusoidal modeling and the periodic signal analysis techniques, by considering periodic signals with non-integer period and global amplitude variation and time warping. And, we show that this model provides a good tradeoff between modeling and estimation errors.

The structured vs. sparse duality considered here is one of parametric representations (in terms of periodic waveform, global amplitude and phase) vs. partially parametric representations (amplitude of fixed atoms). This paper compares the performance of structured vs. sparse representations applied to noise reduction, and underdetermined BSS. A general presentation of the sparse representation of audio signal, and the Matching Pursuit algorithm are proposed in section II. In section III, audio signal extraction based on the global modulation model is described . Finally, the computational complexity, and performance of the two approaches are compared in section IV.

II. Audio Processing with Sparse Representation

A. Time-Frequency atomic decomposition

The decomposition of signals over family of functions that are well localized both in time and frequency has found many applications in signal processing and harmonic analysis. Depending upon the choice of time-frequency atoms, the decomposition might have very different properties.

A general family of time-frequency atoms can be generated by scaling, translating and modulating a single window function $g(t) \in L^2(R)$. For any scale $s > 0$, frequency modulation ξ and translation u, we denote $\gamma = (s, u, \xi)$ and define

$$g_\gamma(t) = \frac{1}{\sqrt{s}} g(\frac{t-u}{s}) e^{i\xi t} \ . \qquad (1)$$

The index γ is an element of the set $\Gamma = R^+ \times R^2$.

The family $D = (g_\gamma(t))_{\gamma\in\Gamma}$ is extremely redundant. To represent efficiently any function $f(t)$, we must select an appropriate countable subset of atoms $(g_{\gamma_n}(t))_{n\in N}$ so that $f(t)$ can be written

$$f(t) = \sum_{n=-\infty}^{+\infty} a_n g_{\gamma_n}(t)$$

Depending upon the choice of the dictionary, the expansion coefficients a_n give explicit information on certain types of properties of

Eurécom's research is partially supported by its industrial partners: BMW, Bouygues Telecom, Cisco Systems, France Télécom, Hitachi Europe, SFR, Sharp, ST Microelectronics, Swisscom, Thales. The research reported herin was also partially sponsored by the french RIAM project SIEPIA

$f(t)$. Windowed transforms (such as Windowed Fourier Transform, Windowed Cosine Transform) and wavelet transforms correspond to different families of time-frequency atoms, that are frames or bases of $L^2(R)$ [3]. In a windowed transform, all the atoms g_{γ_n} have a constant scale $s_n = s_0$ and are thus mainly localized over an interval whose size is proportional to s_0. On the other hand, the wavelet transform decomposes signals over time-frequency atoms of varying scales, called wavelets. A wavelet family $\left(g_{\gamma_n}(t)\right)_{n \in N}$ is built by relating the frequency parameter ξ_n to the scale s_n via $\xi_n = \frac{\xi_0}{s_n}$, where ξ_0 is a constant. The resulting family is composed of dilations and translations of a single function.

B. Matching Pursuit (MP) for audio signal decomposition

Let us consider a family of vectors $D = \left(g_\gamma\right)_{\gamma \in \Gamma}$ included in a Hilbert space H with a unit norm $\|g_\gamma\| = 1$. For a given $f \in H$, getting the best M^{th} order approximant, i.e.

$$\widehat{f}_M = \sum_{m=1}^{M} c_m g_{\gamma_m} = \arg \min_{c_m, \gamma_m} \left\| f - \sum_{m=1}^{M} c_m g_{\gamma_m} \right\| \qquad (2)$$

is an NP-hard problem. The matching pursuit [3] is a greedy strategy to decompose a signal into a linear combination of atoms chosen among the dictionary D. It iteratively defines an m^{th} order residual $R^{m-1}f$ (starting with $R^0 f = f$) in the following way.

1) Compute for all $\gamma \in \Gamma$

$$\left|\left\langle R^{m-1}f, g_\gamma \right\rangle\right|^2 \qquad (3)$$

2) Choose an element $g_{\gamma_m} \in D$ which "closely" matches the residual $R^{m-1}f$ in the sense that

$$\left|\left\langle R^{m-1}f, g_{\gamma_m} \right\rangle\right|^2 = \sup_{\gamma \in \Gamma} \left|\left\langle R^{m-1}f, g_\gamma \right\rangle\right|^2 \qquad (4)$$

3) Compute the new residual by removing the component along the selected atom

$$R^m f = R^{m-1}f - \left\langle R^{m-1}f, g_{\gamma_m} \right\rangle g_{\gamma_m} \qquad (5)$$

The error $\left\|R^M f\right\|$ is proved to decay to zero [4]. Thus, we obtain the atomic decomposition of the signal

$$f = \sum_{m=1}^{\infty} \left\langle R^{m-1}f, g_{\gamma_m} \right\rangle g_{\gamma_m}$$

C. Harmonic Matching Pursuit (HMP) for audio signal decomposition

Gribonval and Bacry propose a variant of the MP algorithm for audio applications [6]. They introduce the harmonic dictionary which extends the Gabor dictionary ($g(t)$ is Gaussian) and better fits the harmonic structure of audio signals. At each step, an atom and all its (approximately) harmonically related atoms get selected. Thanks to the quasiorthogonality of the Gabor atoms $\left(\left\langle g_{\gamma_p}, g_{\gamma_q} \right\rangle \approx \delta\right)$, the Harmonic MP has the same structure as the standard MP, where the inner product is replaced by the correlation function:

$$C\left(R^m, g_{s,u,\xi}\right) = \sum_{k=1}^{K} \sup_{|\xi_k - k\xi|} \left|\left\langle R^m, g_{s,u,\xi_k} \right\rangle\right|^2 \qquad (6)$$

where K is the number of partials in a given harmonic atom.

III. Audio Processing with Structured Representation

A. Signal Model

In sinusoidal modeling, the signal is modeled as a sum of evolving sinusoids:

$$s(t) = \sum_{k=0}^{P} A_k(t) \cos\left(\theta_k(t)\right) \quad . \qquad (7)$$

where $\theta_k(t)$ represents the instantaneous phase of the k^{th} partial. As the music signal is quasi-periodic, $\theta_k(t)$ can be decomposed into

$$\theta_k(t) = 2\pi k t f_0 \; + \; 2\pi \varphi_k(t) \qquad (8)$$

where $\varphi_k(t)$ characterizes the evolution of the instantaneous phases around the k^{th} harmonic; and can be assumed to be lowpass.
The Global Modulation assumption implies that all harmonic amplitudes evolve proportionally in time; and that the instantaneous frequency of each harmonic is proportional to the harmonic index:

$$\begin{cases} A_k(n) = A_k \; A(n) \\ 2\pi\varphi_k(n) = 2\pi k \; \varphi(n) \; + \; \Phi_k \end{cases} \quad . \qquad (9)$$

In summary, we model an audio signal as the superposition of harmonic components with a global amplitude modulation and time warping (that can be interpreted in terms of phase variations):

$$\begin{aligned} & y(n) \;=\; s(n) \;+\; v(n) \\ & =\; \textstyle\sum_k A_k(n) \;\; \cos\left(2\pi k n f_0 + 2\pi\varphi_k(n)\right) + v(n) \\ & =\; A(n) \textstyle\sum_k A_k \cos\left(2\pi k f_0 \left(n + \frac{\varphi(n)}{f_0}\right) + \Phi_k\right) + v(n) \end{aligned}$$

where

- v_n is an additive white Gaussian noise.
- $A(n)$ represents the amplitude modulating signal. It allows an evolution of the note power, reflecting attack, sustain, and decay.
- $\varphi(n)$ denotes the phase modulating signal (that can be interpreted in terms of time warping). The time warping focuses on the time evolution of the instantaneous frequency, and allows the modeling of several musical phenomena (vibrato, glissando...)

In [10], we have expressed the time warping in terms of an interpolation operation over a basic periodic signal. The audio signal can be written as:

$$Y = \underbrace{A \; F\theta}_{= S} \; + V \qquad (10)$$

where :
- $Y = [y(1) \cdots y(N)]^T$, represents the observation vector.
- $S = [s(1) \cdots s(N)]^T$, represents the signal of interest.
- $V = [v(1) \cdots v(N)]^T$, denotes the noise vector.
- $\theta = [\theta(1) \cdots \theta(\lceil T \rceil)]$, characterizes the harmonic signature over essentially one period
- $A \;=\; diag[A(1) \cdots A(N)]$, represents the global amplitude modulation signal.
- F is an $N \times \lceil T \rceil$ interpolation matrix characterizing the time warping. See [10] for a detailed description.

B. Audio Signal Estimation

The previous model is linear in θ, A, or F (separately), F being parameterized nonlinearly. Trying to estimate all factors jointly is a difficult nonlinear problem. Indeed, as the noise is assumed to be a white Gaussian signal, the ML approach leads to the following least-squares problem:

$$\min_{A,F,\theta} \|Y - A\,F\,\theta\|^2 \tag{11}$$

where A and F are parameterized in terms of subsamples. The estimation can easily be performed iteratively though [10]:

1) Periodic Signature Estimation: If we assume that the matrices $\widehat{A}, \widehat{F}$ are given, the periodic signature θ can be isolated as

$$Y = \widehat{A}\,\widehat{F}\,\theta + V = H\,\theta + V \tag{12}$$

Then minimizing (12) w.r.t. θ leads to

$$\widehat{\theta} = \left(H^T H\right)^{-1} H^T Y \quad . \tag{13}$$

Hence the periodic signature gets estimated by using the data over the whole note duration.

2) Instantaneous Amplitude Estimation: In the same manner, the instantaneous amplitude signal can get estimated using a Least-Squares technique [11]. However, we have remarked that the proposed technique is very sensitive to the initialization. Hence, we have proposed to initialize the amplitude estimation based on the estimated powers of the noisy data and the noise.

By assuming the instantaneous amplitude to be piecewise constant, $A(n)$ gets estimated using:

$$\widehat{A}(n) = \sqrt{\frac{1}{\overline{\theta^2}}\left\langle y^2(n) - (y(n) - \widehat{s}(n))^2\right\rangle_n} \tag{14}$$

where $\langle\ .\ \rangle_n$ denotes temporal averaging over the piecewise interval containing n; $\widehat{S} = \widehat{A}\widehat{F}\widehat{\theta}$ denotes the latest estimate of the signal of interest.

3) Instantaneous Frequency Estimation: As for the instantaneous amplitude, the instantaneous frequency gets estimated on a frame-by-frame basis. In each frame, the instantaneous frequency is optimized using (3):

$$\begin{cases} \min\limits_{f} \left\| Y - \widehat{A}\widehat{F}(f)\widehat{\theta} \right\| \\ \frac{\Delta f}{f_0} \leq \alpha_{max} \end{cases} \tag{15}$$

where Δf denotes the maximum relative frequency variation in the current frame compared to the previous frame, reflecting an assumed limited frequency variation rate. The optimal instantaneous frequency value for the current frame gets determined from a finite set of discrete values within the thus limited range.

C. Structured Signal Model vs. Atomic Decomposition

We model an audio signal as a superposition of harmonic components with a global amplitude and frequency modulation:

$$s(n) = A(n) \sum_k A_k \cos\left(2\pi k\left(f_0 n + \varphi(n)\right) + \Phi_k\right)$$

The instantaneous amplitude $(A(n))$, and phase $(\varphi(n))$ are assumed to be lowpass. Hence they can be downsampled. The remaining samples can be estimated using mathematical interpolation, i.e.,

$$A(n) = \sum_p A_p w_A(n - pT_A),\ \varphi(n) = \sum_q \varphi_q w_\varphi(n - qT_\varphi)$$

where $\{A_p, \varphi_q\}$ are the degrees of freedom of the model, (T_A, T_φ) are respectively the downsampling factors of the instantaneous amplitude and phase, and $(w_A(n), w_\varphi(n))$ are given interpolation windows.

If we assume that the instantaneous frequency is piecewise constant ($w_\varphi(n)$ is a triangular window), the quasi-periodic signal model can be interpreted as a sum of a scaled, translated, and modulated harmonic atoms. The basic atom window is given by:

$$g(t) = w_A(n) \sum_k A_k \cos\left(2\pi k n f_0 + \Phi_0\right) = w_A(n)\theta(n) \tag{16}$$

However, the dictionary is not fixed (as in the classic atom decomposition approaches); the atoms are adapted to the signal.

IV. Results

A. Complexity issues

In [3], the authors propose a fast implementation of the Matching Pursuit algorithm based on a structured update of the inner product in (3); and using FFT-based algorithms with appropriate window. They show that the Harmonic Matching Pursuit algorithm has a complexity of $O(KN\log N)$ per iteration (the complexity of the MP algorithm is deduced by making $K = 1$).

On the other hand, the Quasi-Periodic Signal Extraction (QPSE) algorithm can be also be implemented in an efficient way. In fact, interpolation matrices are structured, sparse matrices. Then one can show that the extraction algorithm can be implemented with a $O(NT) + O(T^3)$ complexity per iteration (where T denotes the basic period of the quasi-periodic signal).

B. Noise Reduction

The quality of an audio signal captured in real-world environments is invariably degraded by acoustic interference. This interference can be broadly classified into two distinct categories: additive and convolutive. The noise reduction can be described as the processing of audio signals to reduce the additive noise.

The QPSE algorithm can be applied to extract the audio components of the received signal. If the noise variance $\left(\sigma_v^2\right)$ is available, this information can be used to enhance the estimation of the instantaneous amplitude $\left(\widehat{A}(n) = \sqrt{\frac{1}{\overline{\theta^2}}\left\langle y^2(n) - \sigma_v^2\right\rangle_n}\right)$.

MP-based approaches are also proposed to solve the noise reduction problem [15]. Once the received signal is decomposed into atoms, the noise reduction is performed by classifying the atoms into ”noise” vs. ”signal”; then resynthesizing the enhanced audio signal.

To compare the enhancement accuracy of the two approaches, we have experimented with a real music signal. The proposed signal represents a single note (pitch = 84 Hz) played by an acoustic guitar. The record has a duration of 1s and is sampled at 22.050 Khz. A synthetic Gaussian white noise is added to the audio signal. Furthermore, we consider the global signal-to-noise ratio (SNR_{out}) (possibly limited to the steady-state portion) as an objective evaluation criterion

$$SNR_{out} = 10\log\frac{\sum_{n=1}^{N} s^2(n)}{\sum_{n=1}^{N}\left(s(n) - \widehat{s}(n)\right)^2}$$

which is consistent with previous enhancement studies [13], [14]. Fig. 1 plots curves of the averaged output SNR (evaluated by Monte-Carlo techniques) of Matching Pursuit and Quasi-Periodic Signal Extraction techniques.

We observe that the QPSE and MP approaches have comparable enhancement performance. However, the QPSE approach outperforms the MP in the steady-state region (where the quasi-periodic model allows a better fit of the audio signal). The MP is better in the transition region, where the structure of QPSE is too constrained. We remark also that knowing the noise variance does not significantly increase the enhancement performance for the QPSE approach.

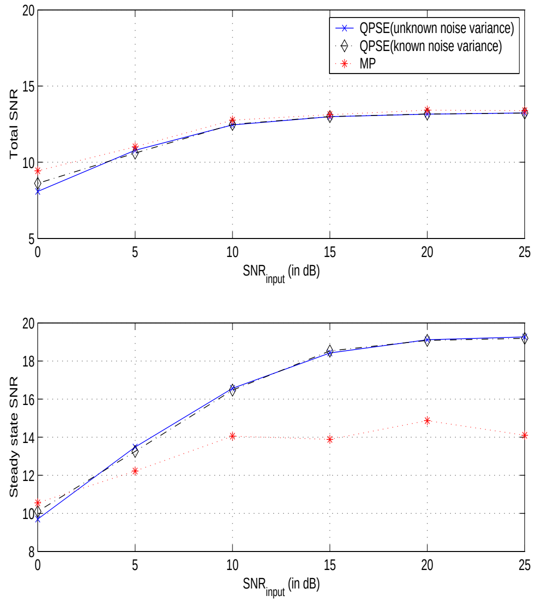

Fig. 1. Noise Reduction (MP on solid line, QPSE on dotted line).

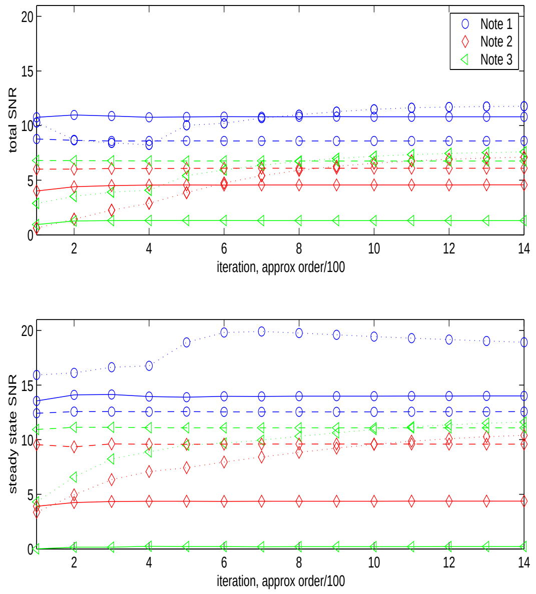

Fig. 2. Separation SNR for mono-mixture audio source separation (MP on solid line, HMP on dashed line, and QPSE on dotted line).

C. *Underdetermined audio source separation*

Blind Source separation is a problem that arises when one or several sensor(s) record data to which can contribute several generating physical processes. It consists in recovering N unknown sources from M instantaneous mixtures. One challenging problem is the underdetermined BSS. It refers to the case when there are less mixtures than sources. In fact, the mixing matrix is not invertible and the traditional BSS methods do not work. And in contrast to the overdetermined case, the source extraction itself requires additional assumptions on the source statistics or structure.

Several authors have proposed underdetermined BSS algorithms that are based on some time-frequency/time-scale representation of the data followed by binary masking [12]. The key observation is that a good data representation often makes it possible to decompose a single underdetermined BSS problem into several (over)determined problems. In the one microphone setting, the underlying hypothesis is that at most one source is ”active” in each component of the representation. The basic principle is simply to :

- decompose the observations into ”components” (atoms).
- perform separation on each atom (which becomes a classification problem).

Another approach is introduced in [11]; where we propose an Iterated Successive Interference Cancellation approach based on the quasi-periodic signal extraction technique.

Using the proposed approaches, we consider separation using a single musical record. The proposed signal represents a synthesized mixture of three notes played by an acoustic guitar. The record has a duration of 1s and is sampled at 22.050 kHz. Their pitch frequencies are respectively 82 Hz, 92 Hz, 116 Hz. Separation SNR (using Matching Pursuit and quasi-periodic signal extraction techniques) are plotted in figure 2. We compare the best separation performance of the different approaches (we focus on the region iteration $\rightarrow \infty$, approximation order $\rightarrow \infty$)

We remark that Matching Pursuit fails to recover the note 3; and that taking into account the harmonic structure of the audio signal increases the separation performance (especially in the steady state region). We see also that the QPSE-based approach outperforms the MP and the HMP approaches, and produces even much better auditive results.

References

[1] F. Abrard, Y. Deville, and P. R. White.“From Blind Source Separation to Blind Source Cancellation in the Underdetermined Case: a new Approach Based on Time-Frequency Analysis,” *In Proc. of ICA*, December 2001.

[2] L-T Nguyen, A. Belouchrani, K. Abed-Meraim, and B. Boashash.“Separating more sources than sensors using Time-Frequency Distributions,” *In Proc. of ISSPA*, August 2001.

[3] S. Mallat and Z. Zhang. “Matching pursuit with time-frequency dictionaries,” *IEEE Trans. on Signal Processing*, Vol.41, pp. 3397-3415, December 1993.

[4] R. Gribonval and P. Vandergheynst. “On the exponential convergence of Matching Pursuits in quasi-incoherent dictionaries,” *Technical report*, IRISA, April 2004.

[5] R. Gribonval, Ph. Depalle, X. Rodet, E. Bacry, and S. Mallat. “Sound Signals Decomposition Using a High Resolution Matching Pursuit,” *In Proc. of ICMC*, 1996.

[6] R. Gribonval, and E. Bacry. “Harmonic Decomposition of Audio Signal with Matching Pursuit,” *IEEE Trans. on Signal Processing*, Vol.51, No.1, January 2003.

[7] S. Krstulovic, R. Gribonval, P. Leveau, and L. Daudet. “A comparison of two extensions of the matching pursuit algorithm for the harmonic decomposition of sounds,” *In Proc. of IEEE WASPAA workshop* , October 2005.

[8] D.D. Muresan, and T.W. Parks. “Orthogonal, Exactly Periodic Supspace Decomposition,” *IEEE Trans. on Signal Processing*,Vol. 51, No. 9, September 2003.

[9] J.D. Wise, J.R. Caprio and T.W. Parks. “Maximum Likelihood pitch estimation,” *IEEE Trans. on Signal Processing* ,Vol. 51, May 1976.

[10] Mahdi Triki, Dirk T.M. Slock. “Periodic Signal Extraction with Global Amplitude and Phase Modulation for Music Signal Decomposition,” *In Proc. of IEEE ICASSP conference*,Vol.3, pp. 233-236, March 2005.

[11] Mahdi Triki, Dirk T.M. Slock. “Multi-Channel Mono-Path Periodic Signal Extraction with Global Amplitude and Phase Modulation for Music and Speech Signal Analysis,” *In Proc. of IEEE SSP workshop*, July 2005.

[12] R. Gribonval. “Piecewise Linear Source Separation,” *In Proc. of SPIE*, Vol.5207, pp. 297-310, August 2003.

[13] J.H.L. Hansen, L.M. Arslan. “Markov model-based phoneme class partitioning for improved constrained iterative speech enhancement,” *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. 3, Issue 1, pp. 98-104, Jan. 1995.

[14] H. Sameti, H. Sheikhzadeh, L. Deng, R.L. Brennan. “HMM-based strategies for enhancement of speech signals embedded in nonstationary noise,” *IEEE Trans. on Signal Processing*, Vol. 6, Issue 5, pp. 445-455, Sept. 1998.

[15] L. Benaroya, F. Bimbot, G. Gravier, and R. Gribonval. “Audio source separation with one sensor for robust speech recognition,” *in ISCA’03*, May 2003.