HAL open science

Underdetermined source separation with structured source priors

Emmanuel Vincent, Xavier Rodet

To cite this version:

Emmanuel Vincent, Xavier Rodet. Underdetermined source separation with structured source priors. 5th Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA), Sep 2004, Granada, Spain. pp.327–332. inria-00544694

HAL Id: inria-00544694

https://inria.hal.science/inria-00544694v1

Submitted on 8 Dec 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Underdetermined Source Separation with Structured Source Priors

Emmanuel Vincent and Xavier Rodet

IRCAM, Analysis-Synthesis Group
1, place Igor Stravinsky
F-75004 PARIS
emmanuel.vincent@ircam.fr

Abstract. We consider the source extraction problem for stereo instantaneous musical mixtures with more than two sources. We prove that usual separation methods based only on spatial diversity have performance limitations when the sources overlap in the time-frequency plane. We propose a new separation scheme combining spatial diversity and structured source priors. We present possible priors based on nonlinear Independent Subspace Analysis (ISA) and Hidden Markov Models (HMM), whose parameters are learnt on solo musical excerpts. We show with an example that they actually improve the separation performance.

1 Introduction

In this article we consider the source extraction problem for stereo instantaneous musical mixtures with more than two sources. The goal is to recover for each sample u the $n \times 1$ vector of source signals $\mathbf{s}_u$ satisfying $\mathbf{x}_u = \mathbf{A}\mathbf{s}_u$, where $\mathbf{A}$ is the $2 \times n$ mixing matrix and $\mathbf{x}_u$ the 2×1 mixture vector. It has been shown that this can be solved in two steps [1]: first estimating the (normalized) columns of $\mathbf{A}$ and then estimating $\mathbf{s}_u$ knowing $\mathbf{A}$. We focus here on this second step.

When little information about the sources is available, the usual hypothesis is that in most time-frequency points only one source is present [2,3,4]. This source is determined exploiting the spatial diversity of the mixture, that is comparing locally the two observed channels. In practice this leads to good results for speech mixtures but not for musical mixtures. Due to western music harmony rules, musical instruments often play notes with overlapping harmonic partials, so that several sources are active in many time-frequency points.

In this article, we investigate the use of structured source priors to improve separation of musical mixtures. We propose a family of priors adapted to instrumental sounds and we show how to use both spatial diversity and source priors into a single separation scheme.

The structure of the article is as follows. In Section 2 we derive a general framework for source extraction and we introduce the three-source example used in the following. In Section 3 we describe some usual separation methods based on spatial diversity and we point their limitations. In Section 4 we propose a family of structured priors adapted to musical sounds and evaluate their performance. We conclude by discussing possible improvements to the proposed method.

2 Source extraction framework

In the rest of the article we suppose that $\mathbf{A}$ has been retrieved from the mixture and has L_2-normalized columns. This is realistic since the spatial directions of the sources can be estimated very precisely when each source is alone in at least one time-frequency point [5]. In this Section we derive a particular piecewise linear separation method and we show that it can potentially recover the sources with very high quality.

2.1 Three-step extraction procedure

Piecewise linear separation methods are three-step procedures [2]: first decompose the mixture channels as weighted sums of time-frequency atoms, then perform a linear separation on each atom, and finally build the estimated sources by summation.

We choose to pass the mixture $\mathbf{x}$ through a bank of filters regularly spaced on the auditory-motivated ERB frequency scale $f_{ERB} = 9.26 \log(0.00437 f_{Hz} + 1)$ to obtain sub-band signals $(\mathbf{x}_f)$. Then we multiply $(\mathbf{x}_f)$ by disjoint 11 ms rectangular windows to compute short-time sub-band signals $(\mathbf{x}_{ft})$. The ERB frequency scale gives more importance to low frequencies which usually contain more energy. This results in a better separation performance than usual linear frequency scales. Note that as a general notation in the following we use bold letter for vectors or matrices, regular letters for scalars and parentheses for sequences.

Because of the linearity of the time-frequency transform, the relationship $\mathbf{x} = \mathbf{A}\mathbf{s}$ becomes $\mathbf{x}_{ft} = \mathbf{A}\mathbf{s}_{ft}$ for each (f, t). A unique solution $\mathbf{s}_{ft}$ can be estimated for each (f, t) by setting some probabilistic priors on the sources. Here we suppose that the source signals $(\mathbf{s}_{jft})$, $1 \leq j \leq n$, are independent and that $(\mathbf{s}_{jft})$ follows a Gaussian prior with known variance m_{jft}. Then the optimal estimated sources are given by $\widehat{\mathbf{s}_{ft}} = \mathbf{\Sigma}_{ft}^{1/2}(\mathbf{A}\mathbf{\Sigma}_{ft}^{1/2})^{+}\mathbf{x}_{ft}$, where $^{+}$ denotes Moore-Penrose pseudo-inversion [1] and $\mathbf{\Sigma}_{ft}$ is the diagonal matrix containing the source variances (m_{jft}). Note that if at least two sources have nonzero variance then perfect reconstruction of the mixture is verified: $\mathbf{x}_{ft} = \mathbf{A}\widehat{\mathbf{s}_{ft}}$.

Finally the waveforms of the estimated sources are obtained by $\widehat{\mathbf{s}} = \sum_{ft} \widehat{\mathbf{s}_{ft}}$.

2.2 Three-source example - Oracle performance

To compare the source extraction methods proposed hereafter, we build an artificial five-second mixture of s_1 = cello, s_2 = clarinet and s_3 = violin, mixed with relative log-powers $\theta_j = \log(A_{2j}^2/A_{1j}^2)$ equal to 4.8 dB, -4.8 dB and 0 dB respectively. In the rest of the article, we separate this mixture with various methods and evaluate the results by computing Source-to-Interference Ratios (SIR) and Source-to-Artifacts Ratios (SAR) [6]. The sources and the mixture are plotted in Fig. 1 and the results are shown in Table 1. All the corresponding sound files can be listened to on the web page `http://www.ircam.fr/anasyn/vincent/ICA04/`.

The first test we make is separation of $\mathbf{x}$ with an oracle estimator of the source power spectrograms $(\mathbf{m}_j)$ (*i.e.* the (m_{jft}) matrices). Performance measures (in

the last line of Table 1) are higher than 20 dB for all sources. This proves that knowing $(\mathbf{m}_j)$ is enough to recover the sources with high quality.

This test mixture is not completely realistic, however it contains instruments sometimes playing in harmony. This results in notes from different instruments overlapping in the time-frequency plane, either partially (during a limited time or on a limited frequency range) or totally. In practice the oracle separation performance cannot be achieved with blind separation methods, because notes that are totally masked cannot be heard and cannot be recovered except with a musical score. However, notes that are partially masked can generally be heard and should be separated accurately.

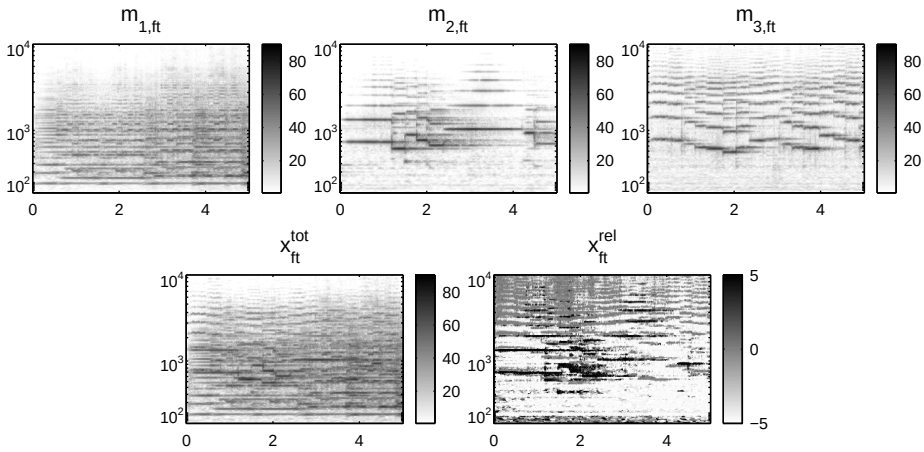

Fig. 1. Power spectrograms of the true sources (top), of the total mixture power and of the relative mixture power (bottom). The horizontal axis is time in seconds, the vertical axis is frequency in Hertz and the color range is in Decibels.

Table 1. Separation of a stereo mixture of three musical sources using several separation methods

Cues	Method	SIR (dB)			SAR (dB)		
		$\widehat{s_1}$	$\widehat{s_2}$	$\widehat{s_3}$	$\widehat{s_1}$	$\widehat{s_2}$	$\widehat{s_3}$
Spatial diversity	All sources	10	8	-5	$+\infty$	$+\infty$	$+\infty$
	Closest source	36	26	18	11.6	10.3	5.6
	1 or 2 closest sources	27	25	15	13.8	13.9	5.9
Source priors	Bernoulli state priors	13	12	6	11.8	10.6	-3.0
Spatial diversity + Source priors	Bernoulli state priors	23	22	34	17.1	16.8	7.0
	Markov state priors	30	31	23	17.2	16.8	8.4
	Oracle state sequence	31	35	23	18.7	18.6	10.5
Oracle	Oracle	49	49	44	24.4	30.0	21.9

3 Separation methods based on spatial diversity

Now that we have explained how to extract the sources given their power spectrograms $(\mathbf{m}_j)$, the problem becomes: how to estimate $(\mathbf{m}_j)$? In this Section we discuss a few heuristic methods based on spatial diversity inspired from [2,3,4,1].

3.1 Some blind separation methods and their performance

Two quantities of interest are computed from the mixture channels $\mathbf{x}_{1,ft}$ and $\mathbf{x}_{2,ft}$: the total log-power $x_{ft}^{\text{tot}} = \log(\|x_{1,ft}\|^2 + \|x_{2,ft}\|^2)$ and the relative log-power $x_{ft}^{\text{rel}} = \log(\|x_{2,ft}\|^2) - \log(\|x_{1,ft}\|^2)$, where we use as synonyms the words "power" and "variance". Heuristic separation methods are based on the following remark: if source j_0 has higher power than the other sources in a given time-frequency point (f,t), then the observed direction x_{ft}^{rel} is close to the direction obtained when only source j_0 is present, that is $\theta_{j_0} = \log(A_{2j_0}^2/A_{1j_0}^2)$.

Suppose without loss of generality that the θ_j are sorted in ascending order. The simplest separation method consists in finding the source j_0 that minimizes $|x_{ft}^{\text{rel}} - \theta_{j_0}|$ and in setting $\widehat{m_{j_0ft}} = 1$ and $\widehat{m_{jft}} = 0$ for $j \neq j_0$: we call this the "closest source" method. A derivation is the "1 or 2 closest sources" method, which is to set $\widehat{m_{1,ft}} = 1$ if $x_{ft}^{\text{rel}} < \theta_1$, $\widehat{m_{n,ft}} = 1$ if $x_{ft}^{\text{rel}} > \theta_n$, and $\widehat{m_{j_0ft}} = 1$ and $\widehat{m_{j_0+1,ft}} = 1$ if $\theta_{j_0} \leq x_{ft}^{\text{rel}} \leq \theta_{j_0+1}$ (and set all other $\widehat{m_{jft}}$ to zero). Finally the "all sources" method consists in setting $\widehat{m_{jft}} = 1$ for all j.

Results for these three separation methods are shown in the first lines of Table 1. Performance is rather good for $\widehat{\mathbf{s}_1}$ and $\widehat{\mathbf{s}_2}$ and lower for $\widehat{\mathbf{s}_3}$, but even for the best method ("1 or 2 closest sources") it remains about 14 dB lower than the oracle performance. There is a compromise between methods that provide high SAR but low SIR ("all sources") and methods that provide high SIR but low SAR ("closest source"). Note that the original "closest source" method described in [3] gave lower performance since it uses only one mixture channel to recover the sources [6]. Computation of mixture sub-bands on a linear frequency scale also yielded lower performance.

3.2 Intrinsic limitation of spatial diversity cues

We generalize these experimental results by showing that spatial diversity cues have intrinsic ambiguities when the sources overlap in the time-frequency plane. When a source $\mathbf{s}_2$ coming from the left ($\theta_{j_2} < 0$) and a source $\mathbf{s}_1$ from the right ($\theta_{j_1} > 0$) are both present in (f,t) with similar powers, then $x_{ft}^{\text{rel}} \approx 0$ so that the source power estimates with the "closest source" method are $\widehat{m_{j_2ft}} = 0$, $\widehat{m_{j_1ft}} = 0$ and $\widehat{m_{j_3ft}} = 1$ for a third source $\mathbf{s}_3$ coming from the center ($\theta_{j_3} \approx 0$). This results in some parts lacking in the "periphery" estimated sources and some excess parts in the "center" estimated sources. This explains why separation performance is generally lower for the "center" source ($\mathbf{s}_3$ here) in a three-source mixture. Note that this limitation generalizes to other estimation methods that use only the single spatial diversity cue x_{ft}^{rel} to determine (m_{jft}). More complex strategies such as [7] suffer from this problem as well in a lesser way.

4 Structured time-frequency source priors

A way to circumvent this limitation is to use the time-frequency structure of the considered sources. Suppose that $\mathbf{s}_1$ and $\mathbf{s}_2$ play notes with harmonic partials. Since instruments play in harmony it is very probable that there exists a time-frequency point (f, t) where $\mathbf{s}_1$ and $\mathbf{s}_2$ have similar power. But if they play different notes at that time or the same note with different spectral envelopes, then it is unprobable that $\mathbf{s}_1$ and $\mathbf{s}_2$ have similar power on all time-frequency points (f', t), $1 \leq f' \leq F$. Using the frequency structure of the sources we can remove the ambiguity in x_{ft}^{rel} using information at all frequencies $\mathbf{x}_t^{\text{rel}} = [x_{1,t}^{\text{rel}}, \ldots, x_{F,t}^{\text{rel}}]^T$. Similarly using the time-structure of the sources we can remove ambiguities when sources are masked locally in time (by percussions for example). A problem remains if $\mathbf{s}_1$ and $\mathbf{s}_2$ have the same power on all frequency range for a large time, since all $\mathbf{x}_t^{\text{rel}}$ provide ambiguous information. This problem may also be circumvented using $\mathbf{x}_t^{\text{tot}}$ in conjunction with $\mathbf{x}_t^{\text{rel}}$. For example if $\mathbf{x}_t^{\text{tot}}$ has energy in high frequency bands only, then it is unprobable that instruments playing only low frequency notes are present at that time.

There are two possibilities to use the time-frequency structure of the sources: either decomposing the mixture on structured time-frequency atoms with priors about the decomposition weights and then using estimation laws of Section 3 to derive $(\widehat{\mathbf{m}}_j)$, or keeping the same time-frequency decomposition as in Section 2 and then deriving $(\widehat{\mathbf{m}}_j)$ with structured priors about $(\mathbf{m}_j)$. We choose here the second solution because musical sources are better described in the time-frequency power domain than in the waveform domain. Relative phases of harmonic partials are rather irrelevant, so that a very large number of atoms would be needed to describe the harmonic structure of most instrumental sounds.

4.1 Structured priors for instrumental sounds

The structured priors we propose here have been used first for single-channel polyphonic music transcription. More details and justifications about our assumptions are available in our companion article [8].

We suppose that each instrument j, $1 \leq j \leq n$, can play a finite number of notes h, $1 \leq h \leq H_j$. At a given time t the presence/absence of note h from instrument j is described with a state variable $E_{jht} \in \{0, 1\}$, and its parameters (instantaneous power, instantaneous frequency, instantaneous spectral envelope, *etc*) are given by a vector of descriptors $\mathbf{p}_{jht} \in \mathbb{R}^{K+1}$. We assume a three-layer generative model, where high-level states (E_{jht}) generate middle-level descriptors $(\mathbf{p}_{jht})$ which in turn generate low-level spectra $(\mathbf{m}_{jt})$. These three layers are termed respectively state layer, descriptor layer and spectral layer.

The spectral layer model is a nonlinear Independent Subspace Analysis (ISA). We write the note descriptors as $\mathbf{p}_{jht} = [e_{jht}, v_{jht}^1, \ldots, v_{jht}^K]$, where e_{jht} is the log-energy of note h from instrument j at time t and (v_{jht}^k) are other variables related to the local spectral shape of this note. Denoting $\mathbf{\Phi'}_{jht}$ the log-power

spectrum of note h from instrument j at time t, we assume

$$\mathbf{m}_{jt} = \sum_{h=1}^{H_j} \exp(\mathbf{\Phi'}_{jht}) \exp(e_{jht}) + \mathbf{n}_j, \tag{1}$$

$$\mathbf{\Phi'}_{jht} = \mathbf{\Phi}_{jh} + \sum_{k=1}^{K} v_{jht}^k \mathbf{U}_{jh}^k, \tag{2}$$

where exp(.) and log(.) are the exponential and logarithm functions applied to each coordinate. The vector $\mathbf{\Phi}_{jh}$ is the total-power-normalized mean log-power spectrum of note h from instrument j and $(\mathbf{U}_{jh}^k)$ are L_2-normalized "variation spectra" that model local variations of the spectral shape of this note. The vector $\mathbf{n}_j$ is the power spectrum of the background noise in source j.

The descriptor layer is defined by setting conditional priors on $\mathbf{p}_{jht}$ given E_{jht}. We assume that e_{jht} is constrained to $-\infty$ and v_{jht}^k to 0 given $E_{jht} = 0$, and that e_{jht} and v_{jht}^k follow independent Gaussian laws given $E_{jht} = 1$.

Finally we consider two models for the state layer in order to study the relative importance of frequential and temporal structure for source separation. A product of Bernoulli priors with constant sparsity factor $P_Z = P(E_{ht} = 0)$ results in frequential structure alone, while a factorial Markov chain prior adds some temporal structure by modeling the typical durations of notes and silences.

4.2 Relationship with the observed mixture

This model for $(\mathbf{m}_j)$ is completed with a model relating $(\mathbf{m}_j)$ to $\mathbf{x}^{\text{tot}}$ and $\mathbf{x}^{\text{rel}}$:

$$\mathbf{x}_t^{\text{tot}} = \log \left[\sum_{j=1}^{n} \mathbf{m}_{jt} \right] + \epsilon_t^{\text{tot}}, \tag{3}$$

$$\mathbf{x}_t^{\text{rel}} = \log \left[\sum_{j=1}^{n} A_{2j}^2 \mathbf{m}_{jt} \right] - \log \left[\sum_{j=1}^{n} A_{1j}^2 \mathbf{m}_{jt} \right] + \epsilon_t^{\text{rel}}. \tag{4}$$

Experiments show that ϵ_t^{tot} and ϵ_t^{rel} can generally be modeled as independent white generalized exponential noises with sparsity parameters $R^{\text{tot}} \simeq 2$ and $R^{\text{rel}} \simeq 0.7$ (*i.e.* ϵ_t^{tot} is Gaussian and ϵ_t^{rel} is sparser than a Laplacian noise).

4.3 Model learning and source power spectra estimation

The probability of $(\mathbf{m}_j)$ given $\mathbf{x}^{\text{tot}}$ and $\mathbf{x}^{\text{rel}}$ is written as the weighted Bayes law

$$P((\mathbf{m}_j)|\mathbf{x}^{\text{tot}}, \mathbf{x}^{\text{rel}}) \propto (P_{\text{spec}})^{w_{\text{spec}}} (P_{\text{desc}})^{w_{\text{desc}}} P_{\text{state}}, \tag{5}$$

involving probability terms $P_{\text{spec}} = \prod_t P(\epsilon_t^{\text{tot}}) P(\epsilon_t^{\text{rel}})$, $P_{\text{desc}} = \prod_{jht} P(\mathbf{p}_{jht}|E_{jht})$ and $P_{\text{state}} = \prod_{jh} P(E_{jh,1}, \ldots, E_{jh,T})$ and correcting exponents w_{spec} and w_{desc}. Weighting by w_{spec} with $0 < w_{\text{spec}} < 1$ mimics the existence of dependencies

between values of ϵ_t^{tot} and ϵ_t^{rel} at adjacent time-frequency points and makes the model distribution closer to the true data distribution.

We learn the model parameters (mean and "variation" spectra, means and variances, initial and transition probabilities) on single-channel solo excerpts of each instrument using a probabilistic model similar to (3) [8].

Then we estimate $(\mathbf{m}_j)$ given $\mathbf{x}^{\text{tot}}$ and $\mathbf{x}^{\text{rel}}$ by finding the states $(\widehat{E_{jht}})$ and the descriptors $(\widehat{e_{jht}})$ and $(\widehat{v_{jht}^k})$ that maximize the posterior (5). Maximization over (E_{jht}) involves a jump procedure with Bernoulli state priors and Viterbi decoding with Markov state priors. Maximization over (e_{jht}) and (v_{jht}^k) is carried out with an approximate second order Newton method. The background noise spectra $(\mathbf{n}_j)$ are re-estimated during transcription to maximize the posterior.

4.4 Performance

The performance of this method was tested using the two defined state models and with an oracle estimator of the state sequence. We also tested separation using only source priors and discarding the spatial likelihood terms $P(\epsilon_t^{\text{rel}})$ in (5). Instrument models were learnt on one-minute solo excerpts taken from other CDs than the test mixture. Results are shown in Table 1.

The combination of spatial diversity and structured source priors provides an average increase of the separation performance of 2.7 dB over spatial diversity alone and 9.7 dB over source priors alone. This proves that our method actually combined the two kinds of information. Results were not significantly improved using larger learning sets.

Moreover results with Markov state priors are a bit better than with Bernoulli state priors, but are still 1.8 dB inferior to results knowing the true state sequence. The main reason for this is not that our method badly estimated the notes played by the instruments, but that some notes were estimated as absent in some zones where they are masked, particularly during reverberation as can be seen in Fig. 2. A way to improve this could be to use more complex state models involving the typical segments "attack, sustain, release" (and reverberation) of musical notes and imposing minimal durations for each of these segments.

5 Conclusion

We considered the source separation problem for underdetermined stereo instantaneous musical mixtures. We proposed a family of probabilistic priors modeling the typical time-frequency structure of musical sources. We showed that combining these priors with spatial diversity leads to a better separation performance than using source priors or spatial diversity alone. This is an important difference with previous works using structured source priors in single-channel [9] and in overdetermined mixtures [10] which did not consider spatial diversity.

A first direction to extend this work is to use simpler source priors involving spectral and temporal continuity but no instrument specific parameters. This could provide faster computations and be useful for the separation of speech

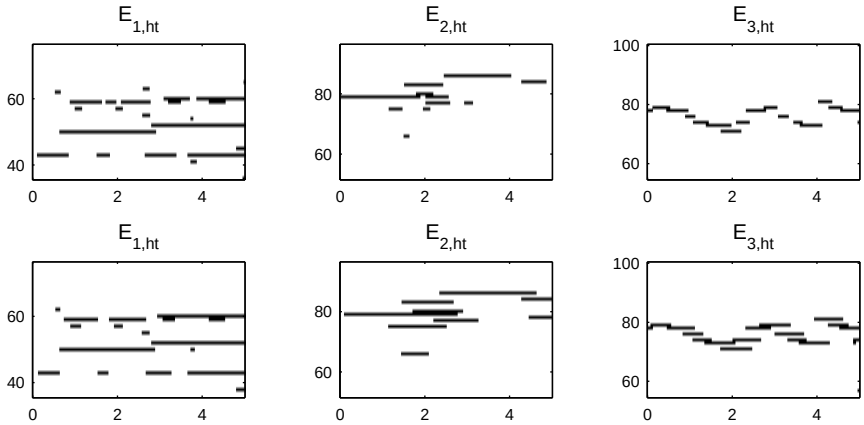

Fig. 2. State sequences obtained with Markov temporal priors (top) compared with oracle state sequences (bottom). The horizontal axis is time in seconds and the vertical axis is note pitch on the MIDI scale.

mixtures. A second direction we are currently considering is to complexify the source priors with other state models, for example forcing instruments to play monophonic phrases, favoring *legato* note transitions or taking into account the "attack, sustain, release" behavior. We are also studying extension of the method to underdetermined stereo convolutive mixtures using other spatial cues.

References

1. Theis, F., Lang, E.: Formalization of the two-step approach to overcomplete BSS. In: Proc. SIP. (2002) 207–212
2. Gribonval, R.: Piecewise linear separation. In: Wavelets: Applications in Signal and Image Processing, Proc. SPIE. (2003)
3. Yilmaz, O., Rickard, S.: Blind separation of speech mixtures via time-frequency masking. IEEE Transactions on Signal Processing (2002) Submitted.
4. Zibulevsky, M., Pearlmutter, B.: Blind source separation by sparse decomposition in a signal dictionnary. Neural Computation **13** (2001)
5. Deville, Y.: Temporal and time-frequency correlation-based blind source separation methods. In: Proc. ICA. (2003) 1059–1064
6. Gribonval, R., Benaroya, L., Vincent, E., Févotte, C.: Proposals for performance measurement in source separation. In: Proc. ICA. (2003)
7. Vielva, L., Erdoğmuş, D., Príncipe, J.: Underdetermined blind source separation using a probabilistic source sparsity model. In: Proc. ICA. (2001) 675–679
8. Vincent, E., Rodet, X.: Music transcription with ISA and HMM. In: Proc. ICA. (2004)
9. Benaroya, L., Bimbot, F.: Wiener based source separation with HMM/GMM using a single sensor. Proc. ICA (2003) 957–961
10. Reyes-Gomez, M., Raj, B., Ellis, D.: Multi-channel source separation by factorial HMMs. In: Proc. ICASSP. (2003)