

# Music Transcription with ISA and HMM

Emmanuel Vincent and Xavier Rodet

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

**Abstract.** We propose a new generative model for polyphonic music based on nonlinear Independent Subspace Analysis (ISA) and factorial Hidden Markov Models (HMM). ISA represents chord spectra as sums of note power spectra and note spectra as sums of instrument-dependent log-power spectra. HMM models note duration. Instrument-dependent parameters are learnt on solo excerpts and used to transcribe musical recordings as collections of notes with time-varying power and other descriptive parameters such as *vibrato*. We prove the relevance of our modeling assumptions by comparing them with true data distributions and by giving satisfying transcriptions of two duo recordings.

## 1 Introduction

In this article we consider the problem of polyphonic music transcription. A musical excerpt can be considered as a time-varying mixture of notes from several musical instruments, where the sound of a given note evolves across time and is described with a set of descriptors (instantaneous power, instantaneous frequency, timbre, *etc*). Given a single-channel musical excerpt and knowing which instruments are playing, we aim at inferring the notes played by each instrument and their descriptors. This can be considered as a semi-blind source separation problem where “meaningful parameters” are extracted instead of waveforms [1]. The main difficulty is that sounds from different instruments are not disjoint in the time-frequency plane and that information about quiet sounds may be masked by louder sounds. Usual approaches are reviewed in [2].

Independent Subspace Analysis (ISA) is a well-suited model for music transcription. Linear ISA describes the short-time power spectrum ( $\mathbf{x}_t$ ) of a musical excerpt as a sum of typical power spectra (or components) ( $\Phi_h$ ) with time-varying weights ( $e_{ht}$ ). This is expressed as  $\mathbf{x}_t = \sum_{h=1}^H e_{ht} \Phi_h + \epsilon_t$  where the modeling error ( $\epsilon_t$ ) is a Gaussian noise [3]. Each note from each instrument is represented by a subspace containing a few components. ISA has been applied to transcription of MIDI-synthesized solo harpsichord [3] and drum tracks [4]. However its robustness to real recording conditions and its ability to discriminate musical instruments have not been studied yet.

The results of existing methods show that linear ISA has three limitations regarding its possible application to the transcription of real musical recordings.

The first limitation is that the modeling error is badly represented as an additive noise term since the absolute value of  $\epsilon_t$  is usually correlated with  $\mathbf{x}_t$ . The modeling error may rather be considered as multiplicative noise (or as additive noise in the log-power domain) [3]. This is confirmed by instrument identification experiments, which use cepstral coefficients (or equivalently log-power spectra) as timbre features instead of power spectra [5,2,6]. The second limitation is that summation of power spectra is not an efficient way of representing the time evolution of note spectra. Many components are needed to represent small fundamental frequency (f0) variations in *vibrato*, wide-band noise during attacks or energy rise of higher harmonics in *forte*. It can easily be seen that summation of log-power spectra is more efficient. The third limitation is that ISA results are not often directly interpretable since many estimated notes with short duration or low power need to be removed before obtaining a readable musical score.

To solve these limitations, we derive here a new nonlinear ISA model considering both summation of power spectra and of log-power spectra and we also study the use of factorial Hidden Markov Models (HMM) as note duration priors.

The structure of the article is as follows. In Section 2 we propose a generative model for polyphonic music combining ISA and HMM. In Section 3 we explain how to learn the model parameters on solo excerpts and how to perform transcriptions. In Section 4 we discuss the relevance of our assumptions and we show two transcription examples. We conclude on possible improvements of the generative model.

## 2 Generative model for polyphonic music

### 2.1 A three-layer generative model

Let  $(\mathbf{x}_t)$  be the short-time log-power spectra of a given polyphonic musical excerpt. As a general notation in the following we use bold letters for vectors, regular letters for scalars and parentheses for sequences. Transcribing  $(\mathbf{x}_t)$  consists in retrieving for each time frame  $t$ , each instrument  $j$  and each note  $h$  from the semitone scale both a discrete state  $E_{jht} \in \{0, 1\}$  denoting presence/absence of the note and a vector of continuous note descriptors  $\mathbf{p}_{jht} \in \mathbb{R}^{K+1}$ . We assume a three-layer probabilistic generative model, where high-level states  $(E_{jht})$  generate middle-level descriptors  $(\mathbf{p}_{jht})$  which in turn generate low-level spectra  $(\mathbf{x}_t)$ . These three layers are termed respectively state layer, descriptor layer and spectral layer. In this Section we describe these layers successively.

### 2.2 Spectral layer with nonlinear ISA

Let us denote  $\mathbf{m}_{jt}$  the power spectrum of instrument  $j$  at time  $t$  and  $\Phi'_{jht}$  the log-power spectrum of note  $h$  from instrument  $j$  at time  $t$ . We write the note descriptors as  $\mathbf{p}_{jht} = [e_{jht}, v_{jht}^1, \dots, v_{jht}^K]$ , where  $e_{jht}$  is the log-energy of note  $h$  from instrument  $j$  at time  $t$  and  $(v_{jht}^k)$  are “variation variables” describing the differences between the mean spectrum of this note and its spectrum at time  $t$ .

Following the discussion of linear ISA limitations in Section 1, we assume

$$\mathbf{x}_t = \log \left[ \sum_{j=1}^n \mathbf{m}_{jt} + \mathbf{n} \right] + \epsilon_t, \quad (1)$$

$$\mathbf{m}_{jt} = \sum_{h=1}^{H_j} \exp(\Phi'_{jht}) \exp(e_{jht}), \quad (2)$$

$$\Phi'_{jht} = \Phi_{jh} + \sum_{k=1}^K v_{jht}^k \mathbf{U}_{jh}^k, \quad (3)$$

where  $\exp(\cdot)$  and  $\log(\cdot)$  are the exponential and logarithm functions applied to each coordinate. The vector  $\Phi_{jh}$  is the total-power-normalized mean log-power spectrum of note  $h$  from instrument  $j$  and the  $\mathbf{U}_{jh}^k$  are  $L_2$ -normalized “variation spectra” related to the “variation variables”. The vector  $\mathbf{n}$  is the power spectrum of the background noise. The modeling error  $\epsilon_t$  is supposed to be a Gaussian white noise with variance  $\sigma_\epsilon^2 \mathbf{I}$ .

Equations (1-2) can be approximated by a simpler nonlinear model using the maximum over each coordinate [7].

### 2.3 Descriptor layer

We assume that note descriptors  $\mathbf{p}_{jht}$  are conditionally independent given the note state  $E_{jht}$ . We set the parametric conditional priors

$$P(\mathbf{p}_{jht} | E_{jht} = 1) = \mathcal{N}_{\mu e_{jh}, \sigma e_{jh}}(e_{jht}) \prod_{k=1}^K \mathcal{N}_{\mu v_{jhk}, \sigma v_{jhk}}(v_{jht}^k), \quad (4)$$

$$P(\mathbf{p}_{jht} | E_{jht} = 0) = \delta_{-\infty}(e_{jht}) \prod_{k=1}^K \delta_0(v_{jht}^k), \quad (5)$$

where  $\mathcal{N}_{\mu, \sigma}(\cdot)$  is the Gaussian distribution of mean  $\mu$  and variance  $\sigma^2$  and  $\delta_\mu(\cdot)$  the Dirac distribution centered in  $\mu$ . For most instruments the parameters  $\mu e_{jh}$ ,  $\sigma e_{jh}$ ,  $\mu v_{jhk}$  and  $\sigma v_{jhk}$  can be shared for all notes  $h$ .

### 2.4 State layer with factorial Markov chains

Finally, we suppose that the states  $E_{jht}$  are independent for different  $(j, h)$ . This results in independence of the descriptors  $\mathbf{p}_{jht}$  for different  $(j, h)$ , which is the usual ISA assumption. We consider two state models: a product of Bernoulli priors and a factorial Markov chain [8], whose equations are respectively

$$P(E_{jh,1}, \dots, E_{jh,T}) = \prod_{t=1}^T (P_Z)^{1-E_{jht}} (1 - P_Z)^{E_{jht}}, \quad (6)$$

$$P(E_{jh,1}, \dots, E_{jh,T}) = P(E_{jh,1}) \prod_{t=2}^T P(E_{jht} | E_{jh,t-1}), \quad (7)$$

where the initial and transition probabilities  $P(E_{jh,1})$  and  $P(E_{jht}|E_{jh,t-1})$  are themselves Bernoulli priors that can be shared for all notes  $h$ .

The silence probability  $P_Z$  is a sparsity factor: the higher it is, the less notes are set as present in a transcription.  $P_Z$  can also be expressed with Markov transition probabilities as  $P_Z = (1 + P(1|0)P(0|1))^{-1}$ . So both models can be used to model the exact sparsity of the transcriptions.

The difference is that the Markov prior adds some temporal structure and favors transcriptions where  $E_{jht}$  is the same on long time segments. HMM have also been used with ISA for source separation in biomedical applications, but in the simpler case of noiseless invertible mixtures [9].

### 3 Learning and transcribing

Now that we have described each layer of the generative model, we explain in this Section how to learn its parameters on solo excerpts and use them to transcribe polyphonic excerpts. We define the instrument model  $\mathcal{M}_j$  as the collection of the fixed parameters related to instrument  $j$ : the spectra  $\Phi_{jh}$  and  $\mathbf{U}_{jh}^k$ , the means and variances  $\mu e_{jh}$ ,  $\sigma e_{jh}^2$ ,  $\mu v_{jhk}$  and  $\sigma v_{jhk}^2$ , and the sparsity factor  $P_Z$ .

#### 3.1 Weighted Bayes

The probability of a transcription  $(E_{jht}, \mathbf{p}_{jht})$  is given by the weighted Bayes law

$$P_{\text{trans}} = P((E_{jht}, \mathbf{p}_{jht}) | (\mathbf{x}_t), (\mathcal{M}_j)) \propto (P_{\text{spec}})^{w_{\text{spec}}} (P_{\text{desc}})^{w_{\text{desc}}} P_{\text{state}}, \quad (8)$$

involving probability terms  $P_{\text{spec}} = \prod_t P(\epsilon_t)$ ,  $P_{\text{desc}} = \prod_{jht} P(\mathbf{p}_{jht} | E_{jht}, \mathcal{M}_j)$  and  $P_{\text{state}} = \prod_{jh} P(E_{jh,1}, \dots, E_{jh,T} | \mathcal{M}_j)$  and correcting exponents  $w_{\text{spec}}$  and  $w_{\text{desc}}$ . Weighting by  $w_{\text{spec}}$  with  $0 < w_{\text{spec}} < 1$  improves the quality of the Gaussian white noise model for  $\epsilon_t$ . This mimics the existence of dependencies between values of  $\epsilon_t$  at adjacent time-frequency points and makes the noise distribution less “peaky” [10].

#### 3.2 Learning and transcription algorithms

Transcribing an excerpt  $(\mathbf{x}_t)$  with instrument models  $(\mathcal{M}_j)$  means maximizing the posterior  $P_{\text{trans}}$  over  $(E_{jht}, \mathbf{p}_{jht})$ .

Transcription with Bernoulli state priors is carried out by reestimating iteratively the note states with a jump procedure. At start all states  $E_{jht}$  are set to 1, then at each iteration at most one note is added or subtracted at each time  $t$  to improve the posterior probability value. Inference with Markov state priors is done with Viterbi decoding. The factorial state space is reduced approximately beforehand by performing inference with a Bernoulli state prior and a low sparsity factor  $P_Z$  to rule out very improbable notes. In both cases the note descriptors are estimated with an approximate second order Newton method.

The modeling error variance  $\sigma_\epsilon$ , the correcting exponents  $w_{\text{spec}}$  and  $w_{\text{desc}}$  and the mean note duration are set by hand to achieve a good compromise between insertion and deletion errors, whereas the power spectrum of the background noise  $\mathbf{n}$  is estimated from the data in order to maximize the posterior.

It is interesting to note that the Newton method updates involve the quantity  $\pi_{jhtf} = \exp(\Phi'_{jhtf}) \exp(e_{jht}) [\sum_{h'=1}^{H_j} \exp(\Phi'_{jh'tf}) \exp(e_{jh't})]^{-1}$  that is the power proportion of note  $h$  into the model spectrum at time-frequency point  $(t, f)$ . When note  $h$  is masked by other notes,  $\pi_{jhtf} \approx 0$  and the value of the observed spectrum  $x_{tf}$  is not taken into account to update  $e_{jht}$  and  $v_{jht}^k$ . This proves that nonlinear ISA performs “missing data” inference [2] in a natural way.

Learning the parameters of a model  $\mathcal{M}_j$  on a solo excerpt ( $\mathbf{x}_t$ ) from instrument  $j$  is done by maximizing the posterior  $P_{\text{trans}}$  iteratively over  $(E_{jht}, \mathbf{p}_{jht})$  and over  $\mathcal{M}_j$ . Models could also be learnt directly on polyphonic excerpts [8], but learning on solo excerpts is more accurate because notes are less likely to be masked.

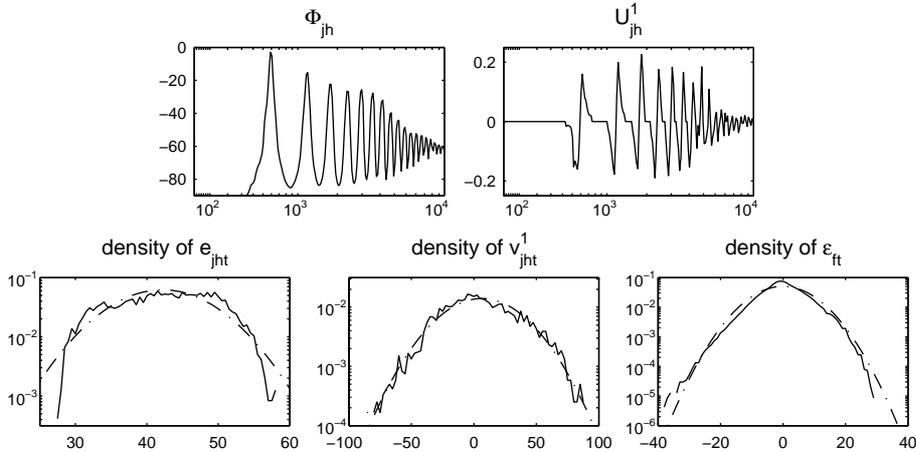
Re-estimation is done again with a Newton method. Note that when  $K > 1$  the spectra  $\mathbf{U}_{jh}^k$  are identifiable only up to a rotation. The size of the model and the initial parameters are fixed by hand. Experimentally we noticed that when the mean spectra  $\Phi_{jh}$  are initialized as spectral peaks at harmonic frequencies they keep this structure during the learning procedure (only the power and the width of the peaks vary). So the learnt spectra really represent “notes” and there is no need of a supervised learning method for the transcriptions to make sense.

### 3.3 Computation of the short-time log-power spectra

The performance of the generative model can be improved by choosing a good time-frequency distribution for  $\mathbf{x}_t$ . In the field of music transcription, nonlinear frequency scales giving more importance to low frequencies are preferable to the linear Fourier scale. Indeed the harmonics of low frequency notes are often masked by harmonics from high frequency notes, so that information in low frequencies is more reliable. Moreover, the modeling of *vibrato* with (3) is relevant only if small f0 variations induce small spectral variations. In the following we use a bank of filters linearly spaced on the auditory-motivated ERB frequency scale  $f_{\text{ERB}} = 9.26 \log(0.00437 f_{\text{Hz}} + 1)$  and we compute log-powers on 11 ms frames (a lower threshold is set to avoid dropdown to  $-\infty$  in silent zones).

## 4 Experiments

Let us now present a few experiments to justify our modeling assumptions and evaluate the performance of the model. We transcribe two real duo recordings: an excerpt from Pachelbel’s canon in D arranged for flute and cello and an excerpt from Ravel’s sonata for violin and cello. We learn instrument models (with  $K = 1$ ) on one-minute solo excerpts taken from other CDs.



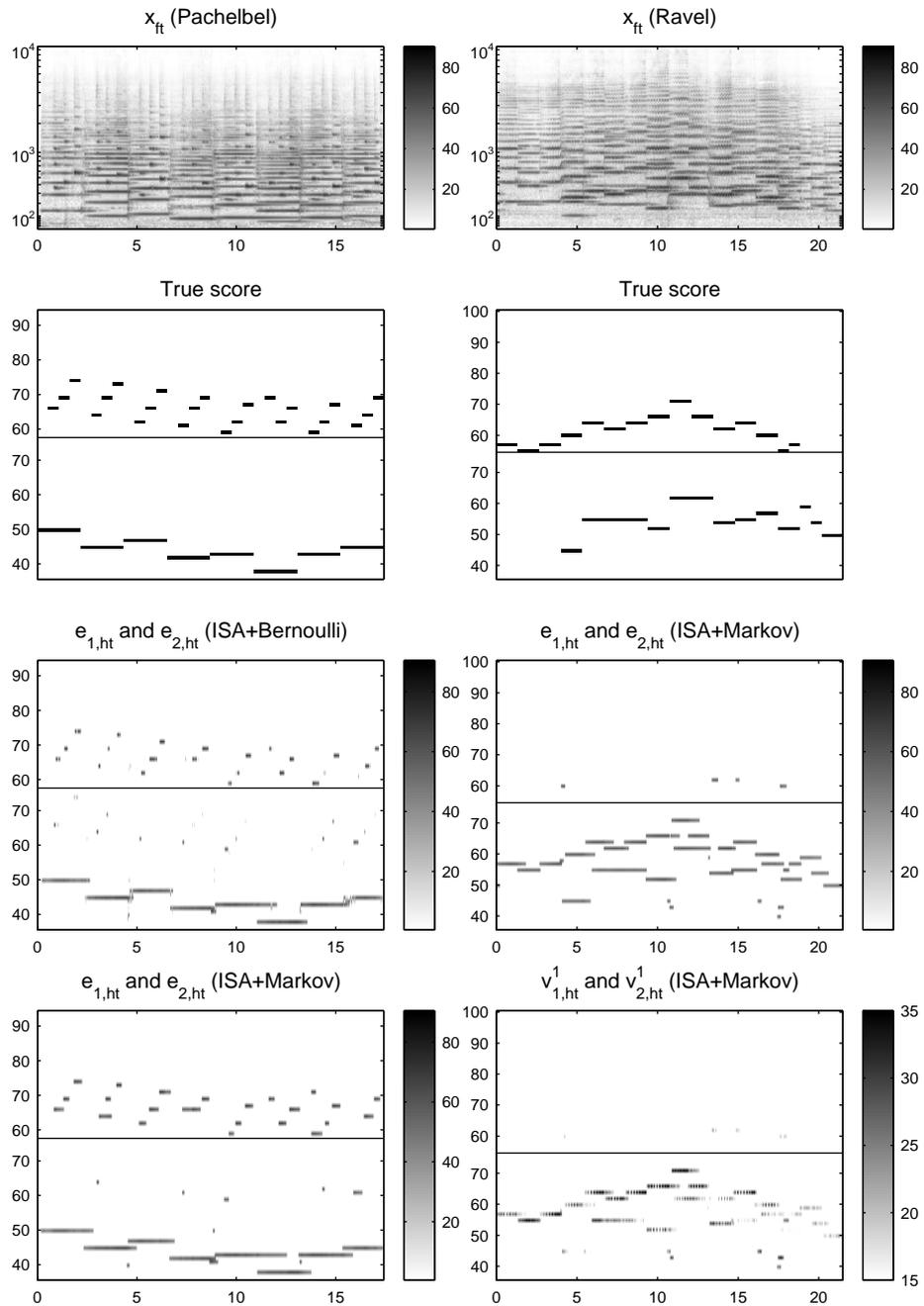
**Fig. 1.** Top: learnt violin spectra  $\Phi_{jh}$  and  $\mathbf{U}_{jh}^1$  with  $h = \text{MIDI } 74$ . Horizontal axis is frequency in Hertz, vertical axis is in Decibels for  $\Phi_{jh}$  and has no unity for  $\mathbf{U}_{jh}^1$ . Bottom: densities of  $e_{jht}$ ,  $v_{jht}^1$  and  $\epsilon_{ft}$  shared for all  $h$  and for  $70 \text{ Hz} \leq f \leq 5000 \text{ Hz}$  (plain line) compared with Gaussians (dash-dotted line). Horizontal axis is in Decibels.

Fig. 1 shows the violin spectra for  $h = \text{MIDI } 74$ . The mean spectrum  $\Phi_{jh}$  contains peaks at harmonic frequencies. And the “variation spectrum”  $\mathbf{U}_{jh}^1$  looks like the derivative versus frequency of  $\Phi_{jh}$  which represents the first order linear approximation of small  $f_0$  variations. Fig. 1 also shows the densities of the variables  $e_{jht}$  and  $v_{jht}^1$  measured on the violin learning excerpt (after transcription with ISA+Bernoulli) and the density of the modeling error  $\epsilon_{ft}$  measured on the first duo (after transcription with ISA+Markov). These densities are all close to Gaussians. We conclude from this that the model actually captured the main spectral characteristics of these musical sounds.

Fig. 2 shows the transcription of the two duos.

The first duo is a difficult example since notes played by the flute are nearly always harmonics of the notes played by the cello, and also belong to the playing range of the cello. The results show that our model is both able to identify most of the notes and to associate them with the right instrument. Results with Markov state model are also a bit better than results with Bernoulli state model, since some spurious short duration notes are removed. There remains 2 note deletions and 10 notes insertions. The inserted notes all have very short durations and could be removed with more complex state models involving rhythm or forcing instruments to play one note at a time (plus reverberation of previous notes).

With the second duo as input, the model is able to identify the notes, but completely fails in associating them with the right instrument. This is not surprising since violin and cello have very close timbral properties, and more complex state models should be used to separate the two note streams. Considering only note transcription and not instrument identification, there are 11 note in-



**Fig. 2.** Transcription of two duo recordings. Top: spectrograms of the recordings. Middle: true scores of the recordings. Below: some estimated note descriptors using various state priors. Horizontal axis is time in seconds, vertical axis is frequency in Hertz (top) or note pitch on the MIDI scale (middle and below). The color range is in Decibels.

sertions that have again very short durations. Note that both instruments play *vibrato*, and that this can be seen in the oscillating values of  $v_{2,ht}^1$ .

## 5 Conclusion

In this article we proposed a new method for polyphonic music transcription based on nonlinear ISA and factorial HMM. This model overcomes the limitations of usual linear ISA by using both summation of power spectra and of log-power spectra and by considering the modeling error as additive noise in the log-power domain. These modeling assumptions were verified on learning data. We also performed satisfying transcriptions of two rather difficult duo recordings.

As noted above, the quality of the transcriptions could be improved by using more complex state models. We are currently considering three directions: constraining instruments to play only one note at a time (plus reverberation of previous notes), using segment models [11] instead of HMM for a better modeling of note durations, and setting temporal continuity priors [4] on the variables  $e_{jht}$  and  $v_{jht}^k$ . We are also investigating the use of instrument models as structured source priors for semi-blind source separation. We present some results in stereo underdetermined mixtures in a companion article [12].

## References

1. Vincent, E., Févotte, C., Gribonval, R.: A tentative typology of audio source separation tasks. In: Proc. ICA. (2003) 715–720
2. Eggink, J., Brown, G.: Application of missing feature theory to the recognition of musical instruments in polyphonic audio. In: Proc. ISMIR. (2003) 125–131
3. Abdallah, S., Plumbley, M.: An ICA approach to automatic music transcription. In: Proc. 114th AES Convention. (2003)
4. Virtanen, T.: Sound source separation using sparse coding with temporal continuity objective. In: Proc. ICMC. (2003)
5. Eronen, A.: Musical instrument recognition using ICA-based transform of features and discriminatively trained HMMs. In: Proc. ISSPA. (2003)
6. Mitianoudis, N., Davies, M.: Intelligent audio source separation using Independent Component Analysis. In: Proc. 112th AES Convention. (2002)
7. Roweis, S.: One microphone source separation. In: Proc. NIPS. (2000) 793–799
8. Ghahramani, Z., Jordan, M.: Factorial hidden Markov models. *Machine Learning* **29** (1997) 245–273
9. Penny, W., Everson, R., Roberts, S.: Hidden Markov Independent Components Analysis. In: *Advances in Independent Component Analysis*. Springer (2000)
10. Hand, D., Yu, K.: Idiot’s bayes - not so stupid after all ? *International Statistical Review* **69** (2001) 385–398
11. Ostendorf, M., Digalakis, V., Kimball, O.: From HMMs to segment models: a unified view of stochastic modeling for speech recognition. *IEEE Trans. on Speech and Audio Processing* **4** (1996) 360–378
12. Vincent, E., Rodet, X.: Underdetermined source separation with structured source priors. In: Proc. ICA. (2004)