Informed source separation using latent components*

Antoine Liutkus, Roland Badeau, Gaël Richard

Institut Telecom, Telecom ParisTech, CNRS LTCI

Abstract. We address the issue of source separation in a particular *informed* configuration where both the sources and the mixtures are assumed to be known during a so-called *encoding* stage. This knowledge enables the computation of a side information which ought to be small enough to be watermarked in the mixtures. At the *decoding* stage, the sources are no longer assumed to be known, only the mixtures and the side information are processed to perform source separation.
The proposed method models the sources jointly using latent variables in a framework close to multichannel nonnegative matrix factorization and models the mixing process as linear filtering. Separation at the decoding stage is done using generalized Wiener filtering of the mixtures. An experimental setup shows that the method gives very satisfying results with mixtures composed of many sources. A study of its performance with respect to the number of latent variables is presented.

1 Introduction

This study concerns a special case of source separation, called *informed source separation* (ISS), that was introduced by Parvaix in [7]. ISS can be understood as an encoding/decoding framework in which both the sources and the mixtures are available at the encoder's side, but only the mixtures are available at the decoder's side, as well as some *side information* that may have been created by the encoder and transmitted along with the mixtures to assist the separation process. ISS thus aims at making source separation robust by providing adequate prior knowledge to the separation algorithms, and allows applications such as *active listening* that consists in being able to mute tracks as in classical Karaoke applications or to add separate effects to them.

The main advantage of ISS is that it permits to reliably recover the separated tracks from the mixtures with only a very small amount of side information. The method we propose here allows to control the quantity of information that is sent to the decoder. As highlighted by Parvaix in [7], if the side information is sufficiently small, it can be directly embedded in the mixture signals by watermarking, allowing active listening of recordings on conventional stereophonic audio CD.

* This work is supported by the French National Research Agency (ANR) as a part of the DReaM project (ANR-09-CORD-006-03) and partly supported by the Quaero Program, funded by OSEO.

We propose a new method for informed source separation that is based on jointly modeling the sources at the encoder side using latent additive independent Gaussian variables in a framework that is very similar to Non-negative Matrix Factorization (NMF). Then, the mixing process is modeled via linear filters that are estimated using techniques inspired from the automatic mixing literature [1]. The *side information* considered in our proposed system thus consists of the spectrum and activation coefficients of each latent component as well as the mixing parameters. The number of necessary bits to store this information is mainly controlled by the number of latent variables considered. At the decoder side, the separation process involves generalized Wiener filtering [2] and allows reaching excellent performance provided that enough latent variables have been chosen.

This paper is organized as follows. In section 2 we detail the model used for representing the source signals as well as the mixing process at the encoder side. In section 3 we describe the estimation method and outline the separation technique induced by the model at the decoder side. Finally, we give some experimental results along with a study of the influence of the number of latent variables on the separation quality in section 4.

2 Model

2.1 Introduction

We consider a set of M source signals $(s_{m,t})_{m=1\cdots M,t=1\cdots L}$ and a set of K mixture signals $(x_{k,t})_{k=1\cdots K,t=1\cdots L}$ of same lengths L. We define $S_{m,\omega n} = [\mathcal{F}(s_{m,\cdot})]_{\omega,n}$ and $X_{k,\omega n} = [\mathcal{F}(x_{k,\cdot})]_{\omega,n}$ as the complex-valued Short Time Fourier Transforms (STFT) of the signals $s_{m,\cdot}$ and $x_{k,\cdot}$ for frequency bin $\omega \in [1 : N_\omega]$ and frame index $n \in [1 : N_n]$.

2.2 Source signals

Complex Gaussian model Following the formalism introduced by Benaroya in [2], the signals of interest are locally modeled as independent wide-sense stationary centered random variables and can thus be characterized by their covariance. In the spectral domain, this can be expressed by writing that the STFT $Y_{\omega n}$ of some signal y_t for frame index n and frequency bin ω obeys[1]:

$$Y_{\omega n} \sim \mathcal{N}_c(0, \sigma^2_{Y_{\omega n}})$$

where $\sigma^2_{Y_{\omega n}}$ is the power spectral density of y_t for the frame n at frequency ω. We further assume that $\{Y_{\omega n}\}_{\omega,n}$ are independent, which stands asymptotically for all (ω, n).

[1] $\mathcal{N}_c$ is the *proper Gaussian complex distribution* and is defined on the plane by its probability density function $f(z) = \frac{1}{\pi\sigma^2} \exp\left(-\frac{|z|^2}{\sigma^2}\right)$

Mixture of latent components The study presented here considers R underlying latent independent centered random Gaussian variables $c_{r,t}$ called the *latent components*, each of which has a power spectral density $W_{\omega r}$ and is only modulated in time by some frame-dependent activation coefficients H_{rn}. In the frequency domain, the STFT $C_{r,\omega n}$ of each latent component $c_{r,t}$ for frame n and frequency ω is thus modeled as:

$$C_{r,\omega n} \sim \mathcal{N}_c(0, W_{\omega r} H_{rn}) \tag{1}$$

Each source signal $s_{m,t}$ is then simply modeled as a weighted sum of these R latent components:

$$s_{m,t} = \sum_{r=1}^{R} \sqrt{Q_{mr}} c_{r,t} \tag{2}$$

where the non-negative coefficient $\sqrt{Q_{mr}}$ is the contribution of the latent component r to source m. As can be seen, all the sources are modeled as a sum of the *same* underlying components. Combining (1) and (2), we have, for frequency bin ω and frame index n:

$$S_{m,\omega n} \sim \mathcal{N}_c(0, \sum_{r=1}^{R} Q_{mr} W_{\omega r} H_{rn}) \tag{3}$$

We can readily see that the model (3) is equivalent to the multichannel NMF model presented in [6]. Indeed, the source signals $s_{m,t}$ are modeled as linear instantaneous mixtures of R latent components. An interesting feature of our model is to allow one single number of latent components for all the source signals.

2.3 Mixing process

Following [1], we will model each mixture signal as a sum of filtered versions of the sources

$$x_{k,t} = \sum_{m=1}^{M} \sum_{\tau=0}^{P} a_{km,\tau} s_{m,t-\tau} = \sum_{m=1}^{M} s_{km,t} \tag{4}$$

where P is the order of the mixing filters, $(a_{km,\tau})_{\tau=0..P}$ is the impulse response of the filter from source m to mixture k and $s_{km,t}$ is called the contribution of source m to mixture k for time t. We will only consider causal and Finite Impulse Response (FIR) filters here. This model can be approximated in the spectral domain as $X_{k,\omega n} = \sum_{m=1}^{M} A_{km,\omega} S_{m,\omega n}$, where $A_{km,\omega}$ is the frequency response of the filter $(a_{km,\tau})_{\tau=0..P}$ at the frequency corresponding to bin ω.

2.4 Unmixing process

During the estimation process, we aim at recovering the original sources $s_{m,t}$ given their contributions $s_{km,t}$ in the mixtures. To that purpose, we follow a

beamforming approach that consists in estimating $s_{m,t}$ as the sum of filtered versions of $s_{km,t}$: $\hat{s}_{m,t} = \sum_{k=1}^{K} \sum_{\tau=0}^{P_u} u_{mk,\tau} s_{km,t-\tau}$, where u_{mk} is the FIR *unmixing filter* of length P_u from mixture k to source m. If U_{mk} is the frequency response of u_{mk}, this can be approximated in the spectral domain as:

$$\hat{S}_{m,\omega n} = \sum_{k=1}^{K} U_{mk,\omega} S_{km,\omega n}. \tag{5}$$

2.5 Set of parameters

The total set Θ of parameters is $\Theta = \{W, H, Q, \mathbf{A}, \mathbf{U}\}$, where $\mathbf{A}$ and $\mathbf{U}$ are respectively composed of all the $M \times K$ impulse responses of the *mixing* and *unmixing* filters from the M sources to the K mixtures and *vice versa*. The total number of parameters is then:

$$\#\Theta = \underbrace{N_\omega \times R}_{\text{for } W} + \underbrace{N_n \times R}_{\text{for } H} + \underbrace{M \times R}_{\text{for } Q} + \underbrace{M \times K \times (P + P_u)}_{\text{for } \mathbf{A} \text{ and } \mathbf{U}} \tag{6}$$

As can be seen from (6), using fixed parameters for the STFT, $\#\Theta$ is controlled by the number R of latent components and the orders P and P_u of the mixing and unmixing filters.

3 Parameters estimation

3.1 Multichannel NMF for source signals

Multiplicative update rules for the parameters For only one source, the model (3) is equivalent to the NMF approach that was popularized by Lee and Seung in [5] when using a particular measure called the Itakura-Saito divergence, which is a special case of β-divergence for $\beta = 0$ (see [4] on this point).

Algorithms in the aforementioned papers can be generalized to the case of M sources and the corresponding update rules for the parameters are summarized in Algorithm 1 for any β-divergence[2].

[2] Notations :

- $.$ denotes element-wise product
- $\frac{A}{B}$ denotes element-wise division
- $M_{m.}$ is the m^{th} row of matrix M
- $[A.^{\alpha}]_{mn} = [A]_{mn}^{\alpha}$
- $\mathbf{S}_m = |S_m|^{.2}$ is the power spectrum of source m
- $\text{diag}(D)$ is a column vector containing the diagonal elements of D if D is a matrix or is the matrix whose diagonal elements are composed of the elements of D if D is a vector.
- $\hat{S}_m = W\text{diag}(Q_{m.})H$ is the estimated power spectrum of source m with current model parameters.

Algorithm 1 Update rules for the parameters of the source model (3) for one iteration

- Q update:
  $Q_{m.} \leftarrow \mathrm{diag}\left(\mathrm{diag}(Q_{m.}).\frac{W^T\left(\hat{S}_m^{\cdot\beta-2}.\mathbf{S}_{m.}(WH)\right)H^T}{W^T\hat{S}_m^{\cdot\beta-1}H^T}\right)$
- W update:
  $W \leftarrow W.\frac{\sum_{m=1}^{M}\left(\hat{S}_m^{\cdot\beta-2}.\mathbf{S}_m\right)(\mathrm{diag}(Q_{m.})H)^T}{\sum_{m=1}^{M}\hat{S}_m^{\cdot\beta-1}(\mathrm{diag}(Q_{m.})H)^T}$
- H update:
  $H \leftarrow H.\frac{\sum_{m=1}^{M}(W\mathrm{diag}(Q_{m.}))^T\left(\hat{S}_m^{\cdot\beta-2}.\mathbf{S}_m\right)}{\sum_{m=1}^{M}(W\mathrm{diag}(Q_{m.}))^T\hat{S}_m^{\cdot\beta-1}}$
- Normalization of W and Q and scaling of H accordingly.

As pointed out by Bertin in [3], better results can be obtained if optimization is first performed with convex cost functions such as the Kullback-Leibler generalized divergence ($\beta = 1$) and then with the Itakura-Saito distance ($\beta = 0$), which is not convex. Such a tempering approach was hence used in this study and indeed proved to show better performance.

3.2 Estimation of the mixing filters

The problem of estimating the mixing filters of different sources in a mixture has already been addressed in so-called *automatic mixing* studies such as [1]. The main idea of these techniques is to choose the mixing filters so as to minimize the mean squared error $\frac{1}{L}\sum_t \left|x_{k,t} - \sum_{m=1}^{M}\left(a_{km} * s_m\right)(t)\right|^2$ for all k. This is done using standard least-squares methods.

3.3 Source separation at the decoder

Sources contributions in mixtures When the parameters of the model have been estimated, we no longer suppose that the source signals $s_{m,t}$ are available. We then focus here on the decoder side, where only the mixtures $x_{k,t}$ and the parameters Θ are available. Considering the mixing model given in 2.3 and the source model (3), we have $S_{km,\omega n} \sim \mathcal{N}_c\left(0, |A_{km,\omega}|^2 \sum_{r=1}^{R} Q_{mr}W_{\omega r}H_{rn}\right)$. If we define $\sigma^2_{km,\omega n} = |A_{km,\omega}|^2 \sum_{r=1}^{R} Q_{mr}W_{\omega r}H_{rn}$, we then have $X_{k,\omega n} \sim \mathcal{N}_c\left(0, \sum_{m=1}^{M}\sigma^2_{km,\omega n}\right)$ and the minimum mean square error (MMSE) estimate of $S_{km,\omega n}$ is thus given by (see also [2,6,4]):

$$\hat{S}_{km,\omega n} = \frac{\sigma^2_{km,\omega n}}{\sum_{m'=1}^{M}\sigma^2_{km',\omega n}} X_{k,\omega n} \tag{7}$$

Sources estimates through beamforming Given all the $\hat{s}_{km,t}$, our objective is now to estimate $s_{m,t}$. Using (5) and (7), we readily see that the estimate $\hat{S}_{m,\omega n}$

of source m for time-frequency bin (ω, n) given Θ and the mixtures is given by:

$$\hat{S}_{m,\omega n} = \sum_{k=1}^{K} U_{mk,\omega} \hat{S}_{km,\omega n} \tag{8}$$

This computation at the decoder side does not require much computational resource as its complexity is $\mathcal{O}(R \times M \times K)$.

The decoder requires the the unmixing filters u_{mk} in order to compute (8). They are included in the parameters set Θ by the encoder, which also computes $\hat{s}_{mk,t}$ following (8) and then chooses u_{mk} so as to minimize the squared error $\frac{1}{L}\sum_t |s_{m,t} - \hat{s}_{m,t}|^2$ for all m.

4 Evaluation

4.1 Corpus and metrics

Corpus Experiments were done with the internal Source Separation Corpus gathered for the Quaero program [3], from which 9 different excerpts were chosen of various musical styles along with their constitutive separated tracks. The corpus includes excerpts constituted of 5 to 11 separated tracks, which are of many kinds, including acoustic instruments such as piano, guitar, male and female singers, distorted sounds/voices, digital effects, etc.

All mixing was done in stereo on real Digital Audio Workstations. It includes equalizing and panning. All sampling rates were set to 44.1kHz and signals are approximately 30s long.

Metrics Objective criteria to evaluate the quality of the separation were used as defined in the *bsseval* toolbox [8] and include the Source to Distortion Ratio (SDR), the Source to Interference Ratio (SIR) and the Source to Artifacts Ratio (SAR). All values are in dB. In order to assess the quality of separation, we have compared the results given by the proposed method to results given by the idealized (oracle) time-frequency mask expressed as follows:

$$\hat{S}_{km,\omega n} = \frac{\|S_{km,\omega n}\|^2}{\sum_{m'=1}^{M} \|S_{km',\omega n}\|^2} X_{k,\omega n}$$

For each excerpt, statistics are averaged over all its constitutive sources in order to give a general overview of the performance of the method. Complete evaluation along with sample signals can be downloaded from our website.

Models parameters All the STFT were computed for frames of 70ms, with 30% overlap. The order of the mixing and unmixing filters a_{km} and u_{mk} were all set to $P = 150$ and the number of iterations for Algorithm 1 was set to 60, the first 30 iterations used $\beta = 1$ and the last 30 iterations used $\beta = 0$. As $\#\Theta$ is mainly controlled by the number R of latent components, we have studied the performance of the method with respect to R.

[3] `www.quaero.org`

4.2 Sources estimates

The results for estimating the sources from the mixtures are given in Figure 1.

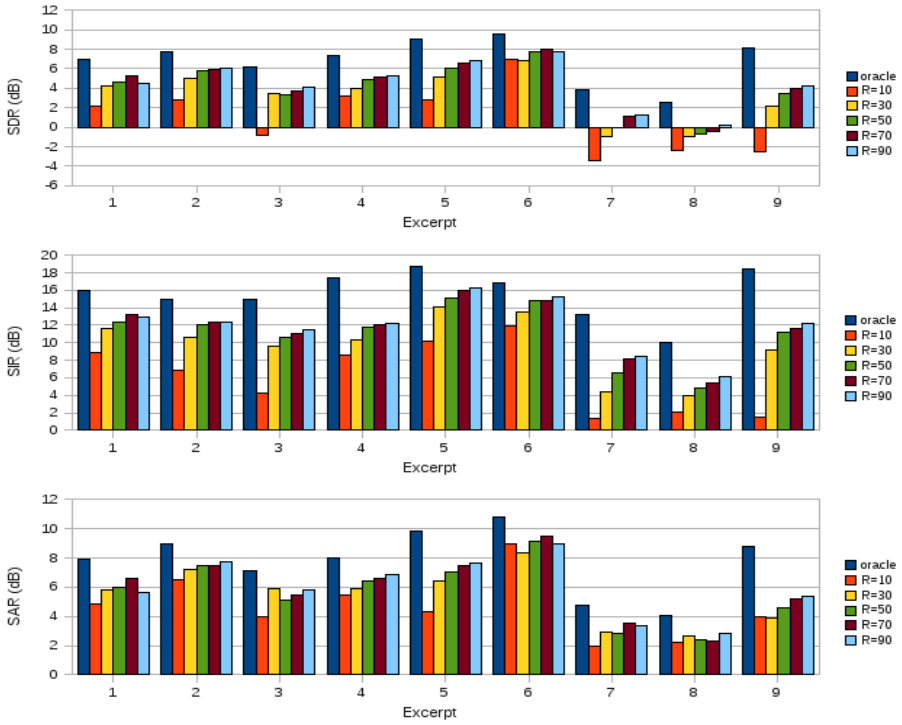

Fig. 1. Average SDR/SIR/SAR scores (in dB) for the estimation of the individual sources. Results are averages over the sources for each excerpt of the corpus.

4.3 Discussion

Several remarks can be made when considering the results given in Figure 1. First, it is perceptually very hard to notice any difference between the original signals and the sources recovered using the oracle method.

Secondly, the quality of the separation is directly controlled by the number R of latent components. As R increases, performance gets closer to the oracle method. There is thus a trade-off between the quality of the separation and the weight of the models parameters. For the results given here, $\#\Theta$ ranges from 1% for $R = 10$ to 7.5% for $R = 90$ of the number of samples in the mixtures.

Finally, even if damaged for small R, the sources are very well isolated one from another, as confirmed by the very high SIR scores.

5 Conclusion

Informed source separation consists in providing valuable prior knowledge to a source separation algorithm. This study considers the case where this knowledge has been computed at an *encoding* stage where both the mixtures and the original sources are known. It then jointly models the source signals through additive latent variables and models the mixing and unmixing processes as linear filters. At the *decoding* stage, separation is performed using generalized Wiener filtering of the mixtures signals.

The total weight of the parameters is extremely small compared to that of the mixtures, typically less than 5 percents. Even though this information is neither quantized nor compressed, this aready allows hiding it directly in the mixture signals through watermarking.

The proposed method allows reaching excellent performance and managed to successfully separate up to 11 sources in stereophonic mixtures during our experiments. The quality of the separation is directly related to the number of latent components used for modeling the sources and can be reliably known by the encoder.

References

1. D. Barchiesi and J. Reiss. Automatic target mixing using least-squares optimization of gains and equalization settings. In *Proc. of the 12th Conf. on Digital Audio Effects (DAFx-09)*, pages 7–14, Como, Italy, September 2009.
2. L. Benaroya, F. Bimbot, and R. Gribonval. Audio source separation with a single sensor. *IEEE Trans. on Audio, Speech and Language Processing*, 14(1):191–199, 2006.
3. N. Bertin, C. Févotte, and R. Badeau. A tempering approach for Itakura-Saito non-negative matrix factorization. With application to music transcription. In *Proc. IEEE Intl. Conf. Acoust. Speech Signal Processing (ICASSP'09)*, pages 1545–1548, Washington, DC, USA, April 2009.
4. C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence. with application to music analysis. *Neural Computation*, 21(3):793–830, 2009.
5. D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, 13:556–562, 2001.
6. A. Ozerov and C. Févotte. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Trans. on Audio, Speech and Language Processing*, 18(3):550–563, 2010.
7. M. Parvaix, L. Girin, and J.-M. Brossier. A watermarking-based method for informed source separation of audio signals with a single sensor. *IEEE Transactions on Audio, Speech and Language Processing*, to be published, 2010.
8. E. Vincent, C. Févotte, and R. Gribonval. Performance measurement in blind audio source separation. *IEEE Trans. on Audio, Speech and Language Processing*, 14(4):1462–1469, 2006.