

# Audio Source Separation Using Hierarchical Phase-Invariant Models

Emmanuel Vincent

METISS Group, IRISA-INRIA  
Campus de Beaulieu, 35042 Rennes Cedex, France  
`emmanuel.vincent@irisa.fr`

**Abstract.** Audio source separation consists of analyzing a given audio recording so as to estimate the signal produced by each sound source for listening or information retrieval purposes. In the last five years, algorithms based on hierarchical phase-invariant models such as single- or multichannel hidden Markov models (HMMs) or nonnegative matrix factorization (NMF) have become popular. In this paper, we provide an overview of these models and discuss their advantages compared to established algorithms such as nongaussianity-based frequency-domain independent component analysis (FDICA) and sparse component analysis (SCA) for the separation of complex mixtures involving many sources or reverberation. We argue how hierarchical phase-invariant modeling could form the basis of future modular source separation systems.

## 1 Introduction

Most audio signals are mixtures of several sound sources which are active simultaneously. For example, speech recordings in “cocktail party” environments are mixtures of several speakers, music CDs are mixtures of musical instruments and singers, and movie soundtracks are mixtures of speech, music and environmental sounds. Audio source separation is the problem of recovering the individual source signals underlying a given mixture.

Two alternative approaches to this problem have emerged: computational auditory scene analysis (CASA) and Bayesian inference. CASA consist of building auditory-motivated sound processing systems composed of four successive modules: front-end auditory representation, low-level primitive grouping, higher-level schema-based grouping and binary time-frequency masking. By contrast, the Bayesian approach consists of building probabilistic generative models of the source signals and estimating them in a minimum mean squared error (MMSE) or maximum a posteriori (MAP) sense from the mixture. The generative models are defined via latent variables and prior conditional distributions between variables. Although individual CASA modules are sometimes amenable to probabilistic models and inference criteria, the Bayesian approach is potentially more robust since all available priors are jointly taken into account via top-down feedback.

Most established Bayesian source separation algorithms rely on time-frequency domain linear modeling [1]. Assuming point sources and low reverberation, the mixing process can be approximated as linear time-invariant filtering. The vector  $\mathbf{X}_{nf}$  of complex-valued short-time Fourier transform (STFT) coefficients of all channels of the mixture signal in time frame  $n$  and frequency bin  $f$  is given by

$$\mathbf{X}_{nf} = \sum_{j=1}^J S_{jnf} \mathbf{A}_{jf} + \mathbf{E}_{nf} \quad (1)$$

where  $S_{jnf}$  are the scalar STFT coefficients of the  $J$  underlying single-channel source signals indexed by  $j$ ,  $\mathbf{A}_{jf}$  are mixing vectors representing the frequency response of the mixing filters and  $\mathbf{E}_{nf}$  is some residual noise. The mixing vectors are typically modeled conditionally to the source directions of arrival via instantaneous or near-anechoic priors, while the source STFT coefficients are modeled as independent and identically distributed according to binary or continuous sparse priors. These priors yield different classes of source separation algorithms, including spatial time-frequency masking, nongaussianity-based frequency-domain independent component analysis (FDICA) and sparse component analysis (SCA) [1].

While these algorithms have achieved astounding results on certain mixtures, their performance significantly degrades on complex sound scenes involving many sources or reverberation [2]. Indeed, due to use of low-informative source priors, separation relies mostly on spatial cues, which are obscured in complex situations. In order to address this issue, additional spectral cues must be exploited. In the framework of linear modeling, this translates into parameterizing each source signal as a linear combination of sound atoms representing for instance individual phonemes or musical notes. In theory, a huge number of sound atoms is needed to obtain an accurate representation since most sources produce phase-invariant atoms characterized by stable variance but somewhat random phase at each frequency. In practice however, only a relatively small number of atoms is usually assumed due to computational constraints, resulting in limited performance improvement [3].

## 2 Hierarchical phase-invariant models

### 2.1 General formulation

Explicit phase invariance can be ensured instead by modeling the source STFT coefficients in a hierarchical fashion via a non-sparse circular distribution whose parameters vary over the time-frequency plane according to some prior. This model appears well suited to audio signals, which are typically non-sparse over small time-frequency regions but non-stationary hence sparse over larger regions. Different distributions have been investigated. Assuming that the source STFT coefficients follow a zero-mean Gaussian distribution, the vector  $\mathbf{X}_{nf}$  of mixture STFT coefficients in time-frequency bin  $(n, f)$  obeys the zero-mean multivariate

Gaussian model [4,5]

$$\mathbf{X}_{nf} \sim \mathcal{N} \left( \mathbf{0}, \sum_{j=1}^J V_{jnf} \mathbf{R}_{jf} \right) \quad (2)$$

where  $V_{jnf}$  are the scalar variances or power spectra of the sources and  $\mathbf{R}_{jf}$  are Hermitian mixing covariance matrices. In the particular case of a stereo mixture, each of these covariance matrices encodes three spatial quantities: interchannel intensity difference (IID), interchannel phase difference (IPD) and interchannel correlation or coherence (IC) [6]. Multichannel log-Gaussian distributions based on these quantities and single-channel log-Gaussian and Poisson distributions have also been proposed [6,7,8].

## 2.2 Prior distributions over the model parameters

Three nested families of prior distributions over the variance parameters  $V_{jnf}$  have been explored so far. In [9,4,5,10,11], the variance of each source is assumed to be locally constant over small time-frequency regions and uniformly or sparsely distributed. In [7,12,13,14], the spectro-temporal distribution of variance is constrained by a Gaussian mixture model (GMM) or, more generally, a hidden Markov model (HMM) that describe each source on each time frame by a latent discrete state indexing one of a set of template spectra. In [15,6,8,16], the spectro-temporal distribution of variance is modeled on each time frame by a linear combination of basis spectra weighted by continuous latent scaling factors. The template spectra and the basis spectra may be either learned on specific training data for each source [7,15,6], or learned on the same set of training data for all sources [12,14] or adapted to the mixture [13,8].

Assuming point sources and low reverberation, the mixing covariance matrices have rank 1 and can be modeled conditionally to the source directions using instantaneous or near-anechoic priors over the aforementioned linear mixing vectors  $\mathbf{A}_{jf}$  as  $\mathbf{R}_{jf} = \mathbf{A}_{jf} \mathbf{A}_{jf}^H$  [9,4]. The model extends to diffuse sources or reverberant conditions, that translate into full-rank mixing covariance matrices. Full-rank uniform priors have been considered in [5,11].

## 2.3 Inference algorithms and results

Approximate inference for the above model is generally carried out by first estimating the model parameters in the MAP sense then deriving the MMSE source STFT coefficients by Wiener filtering. Depending on the chosen priors, different classes of algorithms may be employed to estimate the model parameters, including nonstationarity-based FDICA and SCA [9,10,11], expectation-maximization (EM) decoding of GMM and HMM [7] and nonnegative matrix factorization (NMF) [8]. Although it is not always the most efficient, the EM algorithm is easily applicable in all cases involving a Gaussian distribution.

For single-channel mixtures, the reported signal-to-distortion ratios (SDRs) are on the order of 7 decibels (dB) on mixtures of two speech sources [8] and 10

dB on mixtures of singing voice and musical accompaniment [13]. For stereo mixtures, nonstationarity-based FDICA and SCA have been shown to outperform nongaussianity-based FDICA and SCA by 1 dB or more [17,10,11]. Multichannel NMF has even improved the SDR by 10 dB or more compared to nongaussianity-based FDICA and SCA on certain very reverberant music mixtures with known source directions and learned instrument-specific basis spectra [6].

### 3 Conclusion

To conclude, we believe that hierarchical phase-invariant modeling is a promising framework for research into high-quality source separation. Indeed it allows at the same time accurate modeling of diffuse or reverberant sources and efficient exploitation of spectral cues. These advantages are essential for accurate source discrimination in complex mixtures, involving many sources or strong reverberation. Yet, the potential modeling capabilities offered by this framework have only been little explored. On the one hand, existing models involve a large number of latent variables encoding low-level information. Separation performance and robustness could increase by conditioning these variables on additional latent variables encoding higher-level information, such as reverberation time, source directivity, voiced/unvoiced character and fundamental frequency. On the other hand, most existing source separation systems rely on a single class of priors, which may not be optimal for all sources in a real-world scenario. This limitation could be addressed by designing modular systems combining possibly different classes of priors for each source, selected either manually or automatically. Recent studies in these two directions have introduced promising ideas for future research in this area [18,11,19].

### References

1. Makino, S., Lee, T.W., Sawada, H., eds.: Blind speech separation. Springer (2007)
2. Vincent, E., Araki, S., Bofill, P.: The 2008 Signal Separation Evaluation Campaign: A community-based approach to large-scale evaluation. In: Proc. 8th Int. Conf. on Independent Component Analysis and Signal Separation. (2009) 734–741
3. Gowreesunker, B.V., Tewfik, A.H.: Two improved sparse decomposition methods for blind source separation. In: Proc. 7th Int. Conf. on Independent Component Analysis and Signal Separation. (2007) 365–372
4. Févotte, C., Cardoso, J.F.: Maximum likelihood approach for blind audio source separation using time-frequency Gaussian models. In: Proc. 2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. (2005) 78–81
5. El Chami, Z., Pham, D.T., Servière, C., Guerin, A.: A new model-based underdetermined source separation. Proc. 11th Int. Workshop on Acoustic Echo and Noise Control (2008) paper ID 9061.
6. Vincent, E.: Musical source separation using time-frequency source priors. IEEE Trans. Audio Speech Lang. Process. **14** (2006) 91–98
7. Roweis, S.T.: One microphone source separation. In: Advances in Neural Information Processing Systems 13. (2001) 793–799

8. Virtanen, T., Cemgil, A.T.: Mixtures of gamma priors for non-negative matrix factorization based speech separation. In: Proc. 8th Int. Conf. on Independent Component Analysis and Signal Separation. (2009) 646–653
9. Pham, D.T., Servière, C., Boumaraf, H.: Blind separation of speech mixtures based on nonstationarity. In: Proc. 7th Int. Symp. on Signal Processing and its Applications. (2003) II-73–76
10. Vincent, E., Arberet, S., Gribonval, R.: Underdetermined instantaneous audio source separation via local Gaussian modeling. In: Proc. 8th Int. Conf. on Independent Component Analysis and Signal Separation. (2009) 775–782
11. Duong, N.Q.K., Vincent, E., Gribonval, R.: Spatial covariance models for underdetermined reverberant audio source separation. In: Proceedings of the 2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. (2009)
12. Attias, H.: New EM algorithms for source separation and deconvolution with a microphone array. In: Proc. 2003 IEEE Int. Conf. on Acoustics, Speech and Signal Processing. (2003) V-297–300
13. Ozerov, A., Philippe, P., Bimbot, F., Gribonval, R.: Adaptation of Bayesian models for single-channel source separation and its application to voice/music separation in popular songs. *IEEE Trans. Audio Speech Lang. Process.* **15** (2007) 1564–1578
14. Nix, J., Hohmann, V.: Combined estimation of spectral envelopes and sound source direction of concurrent voices by multidimensional statistical filtering. *IEEE Trans. Audio Speech Lang. Process.* **15** (2007) 995–1008
15. Benaroya, L., McDonagh, L., Bimbot, F., Gribonval, R.: Non negative sparse representation for Wiener based source separation with a single sensor. In: Proc. 2003 IEEE Int. Conf. on Acoustics, Speech and Signal Processing. (2003) VI-613–616
16. Ozerov, A., Févotte, C.: Multichannel nonnegative matrix factorization in convolutive mixtures. With application to blind audio source separation. In: Proc. 2009 IEEE Int. Conf. on Acoustics, Speech and Signal Processing. (2009) 3137–3140
17. Puigt, M., Vincent, E., Deville, Y.: Validity of the independence assumption for the separation of instantaneous and convolutive mixtures of speech and music sources. In: Proc. 8th Int. Conf. on Independent Component Analysis and Signal Separation. (2009) 613–620
18. FitzGerald, D., Cranitch, M., Coyle, E.: Extended nonnegative tensor factorisation models for musical sound source separation. *Computational Intelligence and Neuroscience* (2008) article ID 872425.
19. Blouet, R., Rapaport, G., Cohen, I., Févotte, C.: Evaluation of several strategies for single sensor speech/music separation. In: Proc. 2008 IEEE Int. Conf. on Acoustics, Speech and Signal Processing. (2008) 37–40