Author's final version of article, first published as:
E. Vincent, M. G. Jafari, S. A. Abdallah, M. D. Plumbley, M. E. Davies. Probabilistic modeling paradigms for audio source separation. In W. Wang (Ed), *Machine Audition: Principles, Algorithms and Systems*. Chapter 7, pp. 162-185. IGI Global, 2011. ISBN 978-1-61520-919-4. DOI: 10.4018/978-1-61520-919-4.ch007

Probabilistic Modeling Paradigms for Audio Source Separation

Emmanuel Vincent
IRISA-INRIA, France
Maria G. Jafari, Samer A. Abdallah, Mark D. Plumbley
Queen Mary University of London, United Kingdom
Mike E. Davies
University of Edinburgh, United Kingdom

ABSTRACT
Most sound scenes result from the superposition of several sources, which can be separately perceived and analyzed by human listeners. Source separation aims to provide machine listeners with similar skills by extracting the sounds of individual sources from a given scene. Existing separation systems operate either by emulating the human auditory system or by inferring the parameters of probabilistic sound models. In this chapter, we focus on the latter approach and provide a joint overview of established and recent models, including independent component analysis, local time-frequency models and spectral template-based models. We show that most models are instances of one of the following two general paradigms: linear modeling or variance modeling. We compare the merits of either paradigm and report objective performance figures. We conclude by discussing promising combinations of probabilistic priors and inference algorithms that could form the basis of future state-of-the-art systems.

KEYWORDS
Source separation, latent variable model, spatial model, spectral model, Bayesian inference.

INTRODUCTION
Many everyday sound scenes are produced by several concurrent sound sources: spoken communications are often obscured by background talkers, outdoor recordings feature a variety of environmental sounds, and most music recordings involve a group of several instruments. When facing such scenes, humans are able to perceive and listen to individual sources so as to communicate with other speakers, navigate in a crowded street or memorize the melody of a song (Wang and Brown, 2006). Source separation aims to provide machine listeners with similar skills by extracting the signals of individual sources from a given mixture signal. The estimated source signals may then be either listened to or further processed, giving rise to many potential applications such as speech enhancement for hearing aids, automatic speech and speaker recognition in adverse conditions, automatic indexing of large audio databases, 5.1 rendering of stereo recordings and music post-production.

Depending on the application, the notion of "source" may differ. For instance, musical instruments accompanying a singer may be considered as multiple sources or fused into a single source (Ozerov, Philippe, Bimbot, & Gribonval, 2007). Hence some minimal prior knowledge about the sources is always needed to address the separation task. In certain situations, information such as source positions, speaker identities or musical score may be known and exploited by *informed source separation* systems. In many situations however, only the mixture signal is available and *blind source separation* systems must be employed that do not rely on specific characteristics of the processed scene.

A first approach to audio source separation called computational auditory scene analysis (CASA) is to emulate the human auditory source formation process (Wang and Brown, 2006). Typical CASA systems consist of four processing stages. The signal is first transformed into a time-frequency-lag representation. Individual time-frequency bins are then clustered into small clusters, each associated with one source, by applying primitive auditory grouping and streaming rules. These rules state for example that sinusoidal sounds should be clustered together when they have harmonic frequencies, a smooth spectral envelope, similar onset and offset times, correlated amplitude and frequency modulations, and similar interchannel time and intensity differences. The resulting clusters are further processed using schema-based grouping rules implementing knowledge acquired by learning, such as the timbre of a known speaker or the syntax of a particular language, until a single cluster per source is obtained. The source signals are eventually extracted by associating each time-frequency bin with a single source and inverting the time-frequency transform, an operation known as *binary time-frequency masking*. Although some processing rules may explicitly or implicitly derive from probabilistic priors (Ellis, 2006), the overall process is deterministic: predefined rules implementing complementary knowledge are applied in a fixed precedence order. This bottom-up strategy allows fast processing, but relies on the assumption that each time-frequency bin is dominated by a single source. When this assumption is not satisfied, clustering errors might occur during early processing stages and propagate through subsequent stages.

An alternative approach to source separation is to rely on top-down generative models of the mixture signal that incorporate knowledge about the sound production process. The theory of Bayesian signal processing (Gelman, Carlin, Stern, & Rubin, 2003) provides an appropriate framework to build and exploit such models. The probabilistic distribution of a class of mixture signals is specified by a set of latent variables, including the source signals, and conditional prior distributions between these variables, that are either fixed or learned from training data. For a given mixture signal, the value of any variable may be estimated according to some criterion via standard algorithms such as expectation-maximization (EM) (Dempster, Laird, & Rubin, 1977) or convex or nonconvex optimization (Nocedal & Wright, 2006). Because the auditory system is the product of natural selection based on exposure to natural sounds, the knowledge about the sources exploited by model-based separation systems turns out to be similar to that exploited by CASA (Ellis, 2006). However, contrary to CASA, model-based systems exploit all available knowledge at once and can account for multiple sources per time-frequency bin, hence increasing robustness and maximum potential separation quality. Another advantage of model-based systems is that feedback between high-level and low-level variables allow unsupervised training and adaptation of the systems to unknown data. As a counterpart, computation is generally more intensive.

A large number of model-based audio source separation systems have been presented in the literature in the last twenty years, some of which were individually reviewed elsewhere (Makino, Lee, & Sawada, 2007; Ellis, 2006). Presentations have often focused on historical differences between systems suited to *e.g.* specific numbers of sources and channels, instead of similarities between the underlying models and estimation criteria. In this chapter, we provide an overview of state-of-the-art systems amenable to fully probabilistic generative models and Bayesian estimation criteria, with particular emphasis on recent systems. We identify six classes of models and interpret them as instances of two general paradigms: linear modeling or variance modeling. For each class, we present some popular prior distributions and estimation criteria. We also report typical performance figures and discuss the outcomes of a recent evaluation campaign. We conclude by summarizing the common principles behind most models and the merits of either paradigm and by discussing promising combinations of probabilistic priors and inference algorithms that could form the basis of future source separation systems.

The following notations are employed throughout the chapter. In the general case when the mixture may consist of several channels, multichannel variables are denoted by bold letters and single-channel variables by regular letters. Subscripts represent nested subsets of entries of multidimensional variables, for instance $\mathbf{C}=(\mathbf{C}_1,\ldots,\mathbf{C}_j,\ldots,\mathbf{C}_J)$ and $\mathbf{C}_j=(\mathbf{C}_{j1},\ldots,\mathbf{C}_{jt},\ldots,\mathbf{C}_{jT})$. Calligraphic letters denote standard parametric probability distributions, whose definitions are recalled in appendix.

SOURCE SEPARATION VIA LINEAR MODELING

General principles

The general Bayesian approach to signal processing is to build a generative model of the observed signal based on the available knowledge about its production process. In the context of audio, mixture signals may result from different acquisition techniques. Hearing aid recordings and noisy phone conversations are obtained by simultaneously recording all sources with one or more microphones. Due to the linearity of air propagation, the signal recorded at each microphone is equal to the sum of individual *source components*, which would have been recorded if each source alone was present. Pop music CDs and movie soundtracks are often made not from such live recordings but by synthetic mixing of source signals separately recorded in a studio. The mixing process then consists of applying sound effects to each source signal and adding the resulting multichannel components. In the following, we denote by I the number of channels and by J the number of sources. Regardless of the acquisition technique, the mixture signal $\mathbf{x}_t$ at time t is given by the linear model (Cardoso, 1998)

$$\mathbf{x}_t = \sum_{j=1}^{J} \mathbf{c}_{jt} + \boldsymbol{\varepsilon}_t. \tag{1.1}$$

where $\mathbf{c}_{jt}$ is the jth source component at that time and $\boldsymbol{\varepsilon}_t$ some residual measurement or background noise.

In order to complete the above model, the source components and noise must be characterized by prior distributions. A general principle behind the choice of these priors is that they should provide sufficient knowledge to solve the problem, but not too much. When the chosen priors are too weak, they may *underfit* the characteristics of audio sources and result in poor discrimination. This would happen for instance if all source components and noise followed a uniform prior, in which case any set of source components would be a solution. When the priors are too strong, they may *overfit* the characteristics of the source signals for which they have been designed but badly generalize to other sources.

In the framework of linear modeling, all sources are generally assumed to be *point sources* localized at well-defined, possibly moving, spatial positions. This assumption is valid for speakers and small musical instruments, but not for diffuse or semi-diffuse sources. The effects of air propagation between all sources and microphones or the applied sound effects are then modeled as linear time-varying mixing filters, whose length depends on the amount of reverberation. This filtering process is typically represented by transforming the signals into the time-frequency domain using the short time Fourier transform (STFT), although auditory-motivated or adaptive time-frequency transforms are also applicable (Roman & Wang, 2008; Nesbit, Vincent, & Plumbley, 2009).

Denoting by $\mathbf{X}_{nf}$, $\mathbf{C}_{jnf}$ and $\mathbf{E}_{nf}$ the vectors of complex-valued STFT coefficients of the mixture, the jth source and the residual over all channels in time frame n and frequency bin f, the model becomes

$$\mathbf{X}_{nf} = \sum_{j=1}^{J} \mathbf{C}_{jnf} + \mathbf{E}_{nf}. \tag{1.2}$$

Assuming that the residual noise is Gaussian with covariance $\boldsymbol{\Sigma}^{\mathbf{E}}$, the likelihood of the mixture STFT coefficients $\mathbf{X}$ given the STFT coefficients of the source components $\mathbf{C}$ is equal to

$$P\left(\mathbf{X} \mid \mathbf{C}, \boldsymbol{\Sigma}^{\mathbf{E}}\right) = \prod_{nf} \mathcal{N}\left(\mathbf{X}_{nf}; \sum_{j=1}^{J} \mathbf{C}_{jnf}, \boldsymbol{\Sigma}^{\mathbf{E}}\right). \tag{1.3}$$

Furthermore, under low reverberation conditions, filtering translates approximately into complex-valued multiplication in the Fourier domain, so that each source component is equal to the product of single-channel source STFT coefficients S_{jnf} by some *mixing vectors* $\mathbf{A}_{jnf}$ encoding interchannel intensity and phase differences

$$\mathbf{C}_{jnf} = S_{jnf} \mathbf{A}_{jnf}. \tag{1.4}$$

The I x J matrix $\mathbf{A}_{nf}$ whose jth column is given by the mixing vector $\mathbf{A}_{jnf}$ is called the *mixing matrix*. Both the source coefficients and the mixing vectors are generally unknown and possibly depend on a set of additional latent variables z. The choice of a model consists of identifying relevant variables and setting suitable priors $P(\mathbf{A}|z)$, $P(S|z)$ and $P(z)$. When the same priors are set over all sources, source separation is feasible at best up to permutation indeterminacy.

For a given mixture signal, the STFT coefficients of the source components[i] can be inferred using one of several alternative Bayesian estimation criteria (Gelman et al., 2003) such as minimum mean square error (MMSE)

$$\widehat{\mathbf{C}}_{\text{MMSE}} = \int \mathbf{C} P\left(\mathbf{C} \mid \mathbf{X}\right) d\mathbf{C} \tag{1.5}$$

or maximum a posteriori (MAP)

$$\widehat{\mathbf{C}}_{\text{MAP}} = \arg\max_{\mathbf{C}} P\left(\mathbf{C} \mid \mathbf{X}\right) \tag{1.6}$$

where the posterior distribution of the source components is defined up to a multiplicative constant as

$$P\left(\mathbf{C} \mid \mathbf{X}\right) \propto \iint P\left(\mathbf{X} \mid \mathbf{C}, \boldsymbol{\Sigma}^{\mathbf{E}}\right) P\left(\mathbf{A} \mid z\right) P\left(S \mid z\right) P\left(z\right) P\left(\boldsymbol{\Sigma}^{\mathbf{E}}\right) d\boldsymbol{\Sigma}^{\mathbf{E}} dz. \tag{1.7}$$

Time-domain source components are then computed by inverting the STFT. The MAP criterion is also called maximum likelihood (ML) when the priors are deterministic or uniform. The choice of a criterion depends on the target tradeoff between robustness and accuracy and on computational constraints. In practice, the above integrals are often intractable and approximate inference criteria such as joint MAP estimation of several variables are employed instead.

Binary local time-frequency linear models

The simplest approach to source separation is perhaps to assume that the sources have arbitrary spectro-temporal characteristics but that they are located at different spatial positions. Under this assumption, the source coefficients S_{jnf} in different time-frequency bins can be modeled as independent. Looking at the STFT coefficients of speech signals in Figure 1, it appears that these are *sparse*: few coefficients have significant values and most are close to zero. Moreover, the most significant coefficients of each source are often in different time-frequency bins than for other sources. In the limit where the coefficients are very sparse, it can be assumed that a single source indexed by m_{nf} is active in each time-frequency bin (Yılmaz & Rickard, 2004). With a uniform prior over the coefficient of the active source, the STFT coefficients S_{jnf} of all sources follow the binary local time-frequency model

$$P(m_{nf}) = \mathcal{U}(m_{nf}) \text{ and } \begin{cases} P(S_{jnf} \mid m_{nf}) = \mathcal{U}(S_{jnf}) \text{ if } j = m_{nf} \\ S_{jnf} = 0 \text{ otherwise.} \end{cases} \tag{1.8}$$

In order to achieve source separation with this model, dependencies between the mixing vectors of a given source in different time-frequency bins must be introduced via some prior $P(\mathbf{A}_j)$ (Sawada, Araki, Mukai, & Makino, 2007). For instance, $\mathbf{A}_{jnf}$ is constant over time for nonmoving sources and over both time and frequency for *instantaneous* mixtures generated via amplitude panning. For *anechoic* mixtures recorded in an environment without sound reflections or simulating such an environment, $\mathbf{A}_{jnf}$ is given by

$$\mathbf{A}_{jnf} = \begin{pmatrix} g_{1jn} \mathrm{e}^{-2\mathrm{i}\pi f \tau_{1jn}} \\ \vdots \\ g_{1jn} \mathrm{e}^{-2\mathrm{i}\pi f \tau_{1jn}} \end{pmatrix} \tag{1.9}$$

where g_{ijn} and τ_{ijn} denote respectively the attenuations and the delays associated with direct propagation from source j to all microphones i, which can be determined up to an overall attenuation and delay given the source direction of arrival (DOA) θ_{jn} (Yılmaz & Rickard, 2004). In *echoic* mixtures, reflected sounds result in smearing of the mixing vectors around the above anechoic model. This smearing effect may be either considered as part of the residual noise (Izumi, Ono, & Sagayama, 2007) or modeled via some prior $P(\mathbf{A}_{jnf} \mid \theta_{jn})$ (Mandel, Ellis, & Jebara, 2007; Sawada et al., 2007; Roman & Wang, 2008). Parametric priors $P(\theta_j)$ modeling source movements were also studied in (Roman & Wang, 2008).

Under a noiseless echoic model, joint MAP estimation of the source DOAs and the source components amounts to alternating optimization of the source DOAs and the active source indexes and results in a form of binary time-frequency masking: the STFT coefficients of the active source component in each time-frequency bin are set to those of the mixture, while the coefficients of other sources are set to zero.

This is the principle behind the popular degenerate unmixing estimation technique (DUET) (Yılmaz & Rickard, 2004). MMSE estimation of the source components results in *soft time-frequency masking* instead (Mandel et al., 2007): the estimated STFT coefficients of each source component are equal to the mixture STFT coefficients scaled by the posterior probability of activity of that source, which is a scalar between 0 and 1 determined via EM. Similar estimates are obtained under the noisy anechoic model, except that each source component is projected over the corresponding mixing vector (Izumi et al., 2007). Figure 1 illustrates the application of the latter model to a stereo instantaneous mixture of three speech sources. The clustering of incoming sound directions around the source DOAs is clearly visible. Although the estimated source components exhibit similar features to the original ones, sinusoidal partial tracks are often interrupted and zeroed out.

Source separation systems based binary local time-frequency linear models have been applied to various speech and music mixtures, including instantaneous and anechoic mixtures of nonmoving sources (O'Grady & Pearlmutter, 2008; Yılmaz & Rickard, 2004), live recordings of nonmoving sources (Izumi et al., 2007; Mandel et al., 2007; Sawada et al., 2007) and, more rarely, live recordings of moving sources (Roman & Wang, 2008). Although these systems can recover the source components with a typical signal-to-distortion ratio (SDR) (Vincent, Araki, & Bofill, 2009a) around 10 decibels (dB) for stereo instantaneous or anechoic mixtures of three sources, they generally produce *musical noise* artifacts due to discontinuities introduced between neighboring time-frequency bins. Moreover, performance degrades with additional sources or reverberation, since the number of significant sources per time-frequency bin and the smearing of the mixing vectors around the source DOAs increase.

Continuous local time-frequency linear models

A common critique of the above binary time-frequency models is that they assume a priori dependency between the STFT coefficients of all sources within each time-frequency bin, while several experiments have shown that these coefficients can be considered as independent in most situations (Puigt, Vincent, & Deville, 2009). These coefficients can therefore be modeled in a more principled fashion as independent continuous nongaussian variables (Cardoso, 2001). The sparsity property of audio signals suggests the use of a sparse distribution $P(S_{jnf})$ characterized by a peak at zero and heavy tails with respect to a Gaussian. A common choice is the circular *generalized exponential distribution*

$$P(S_{jnf}) = \mathcal{G}(S_{jnf}; \beta, p) \tag{1.10}$$

where β is a variance-related parameter and p a shape parameter. Circularity means that the distribution has zero mean and is phase-invariant. The shape parameter encodes prior knowledge about sparsity: the smaller p, the sparser the distribution. Figure 2 shows that the generalized exponential closely follows the empirical distribution of $|S_{jnf}|$ for speech sources and that $p \approx 0.4$ for all sources in this example. A range of values of p were explored in (Vincent, 2007). The value $p=1$ yielding the *Laplacian distribution* is often assumed (Winter, Kellermann, Sawada, & Makino, 2007). Alternative circular and non-circular sparse distributions, some of which were implicitly specified by their score function $\Phi(S_{jnf}) = \partial \log P(S_{jnf}) / \partial S_{jnf}$, were used in (Smaragdis, 1998; Zibulevsky, Pearlmutter, Bofill, & Kisilev, 2001; Cemgil, Févotte, & Godsill, 2007). The priors over the mixing vectors $\mathbf{A}_{jnf}$ employed in this context are similar to above and based either on instantaneous mixing assumptions (Zibulevsky et al., 2001), on latent source DOAs (Sawada et al., 2007) or on continuity over frequency (Nesta, Omologo, & Svaizer, 2008).

In general, joint MAP estimation of the source components and other model parameters is achieved via nonlinear optimization. A case of particular interest is when the number of sources is equal to the number of mixture channels and no measurement noise is assumed. In that case, it is sufficient to estimate the mixing vectors $\mathbf{A}_{jnf}$ since the vector of source coefficients can be derived as the inverse of the mixing matrix $\mathbf{A}_{nf}$ multiplied by the vector of mixture coefficients $\mathbf{X}_{nf}$. The inverse mixing matrix acts as a linear spatial filter or *beamformer* that attenuates sounds from certain directions (Trees, 2002). A range of nonlinear optimization algorithms suited to that case were proposed under the name of nongaussianity-based *frequency-domain independent component analysis* (FDICA). Most FDICA algorithms achieve approximate MAP inference by dropping across-frequency dependencies between mixing vectors in a

first stage so as to estimate the source components up to an arbitrary permutation within each frequency bin and finding the optimal permutations in a second stage by exploiting these dependencies (Sawada et al., 2007; Lee, 2009). In the case when the number of sources is larger than the number of mixture channels, *sparse component analysis* (SCA) algorithms may be used instead. Most SCA algorithms also adopt approximate inference strategies, such as deriving the MAP source coefficients given fixed mixing vectors estimated via a binary model (Zibulevsky et al., 2001; Winter et al., 2007). When p is small, MAP inference under a generalized Gaussian prior splits the mixture in each time-frequency bin between the I sources whose DOAs are closest to the incoming sound direction and sets other source components to zero (Vincent, 2007). Figure 2 depicts an example of application of SCA with a Laplacian prior to the stereo mixture of three speech sources previously considered in Figure 1. Comparison between the figures shows that the estimated source components are more accurate with SCA than with binary masking.

FDICA-based systems have been applied to the separation of live recordings of nonmoving sources (Smaragdis, 1998; Sawada et al., 2007; Nesta, Omologo, & Svaizer, 2008) and moving sources (Mukai, Sawada, Araki, & Makino, 2004), and SCA-based systems to the separation of instantaneous or echoic mixtures (Zibulevsky et al., 2001; Peterson & Kadambe, 2003; Cemgil et al., 2007; Vincent, 2007; Nesbit et al., 2009) and live recordings of nonmoving sources (Winter et al., 2007). These systems generally outperform binary or soft masking-based systems. For instance, they achieve a typical SDR around 15 dB for stereo live recordings of two speech sources in a room with moderate reverberation time or for stereo instantaneous mixtures of three speech sources. In addition, FDICA and IVA do not generate musical noise. Nevertheless, performance decreases in reverberant conditions for the same reason as above, namely that the smearing of the mixing vectors around the source DOAs increases.

Linear models over arbitrary basis signals

One approach to improve separation performance is to exploit the spectral characteristics of each source in addition to its spatial characteristics. This approach also makes it feasible to separate single-channel signals. Speech and music involve wideband periodic, near-periodic, noisy or transient sounds. The sounds produced by each source exhibit resonances at specific frequencies. Also, male and female speakers or instruments from the same family usually have different fundamental frequency ranges. Assuming that a given source can produce a few such sounds at a time, each source signal can be modeled in the time domain as the linear combination of a set of wideband basis signals b_{jkt} multiplied by scale factors α_{jk} following independent continuous sparse priors. This model can be equivalently written in the time-frequency domain as

$$S_{jnf} = \sum_{k=1}^{K} \alpha_{jk} B_{jknf}. \tag{1.11}$$

Since a given sound may be produced at any time instant, the set of basis signals must be translation-invariant, *i.e.* any delayed basis signal must also be a basis signal (Blumensath & Davies, 2006). This implies that the number K of basis signals be much larger than the number of samples of the mixture signal. Due to computational constraints, approximate translation invariance is generally assumed instead by constraining each basis signal to be nonzero over a single time frame and restricting the set of possible delays to an integer number of time frames (Jang & Lee, 2004).

This approach has led to a few different inference strategies. In (Jang & Lee, 2004), single-channel mixtures are separated by MAP estimation of the scale factors given fixed basis signals learned in the MAP sense from a different set of training signals for each source. In (Gowreesunker & Tewfik, 2007), the separation of synthetic instantaneous mixtures is achieved by first estimating the mixing vectors using a binary time-frequency model, then deriving the MAP scale factors associated with basis signals learned from a different set of training signals for each source. These two strategies rely on prior knowledge that the sources to be separated are similar to those in the training set. By contrast, in (Blumensath & Davies, 2006), both the basis signals and the scale factors are blindly inferred from the observed single-channel mixture and partitioned into sources using prior dependencies between the scale factors of each source.

The source separation system in (Gowreesunker & Tewfik, 2007) has been reported to provide modest SDR improvement on the order of 1 dB compared to SCA on stereo instantaneous mixtures of three speech sources using a few thousand basis signals per time frame. This limited improvement may be explained by the fact that the representation of spectral characteristics via linear modeling is efficient only for a few sound sources such as electronic musical instruments which generate reproducible waveforms. Most sound sources, in particular near-periodic or noise sources, generate sounds whose relative phase across frequency varies almost randomly from one instance to another. In order to accurately represent such sources, the set of basis signals must be phase-invariant, which greatly increases the number of basis signals and makes inference even less tractable.

SOURCE SEPARATION VIA VARIANCE MODELING

General principles

The search for efficient ways of representing spectral sound characteristics has led to the investigation of an alternative paradigm consisting of modeling the STFT coefficients of each source component by a circular distribution whose parameters vary over time and frequency, thus exploiting nonstationarity and nonwhiteness instead of or in addition to nongaussianity (Cardoso, 2001). In a single-channel context, this distribution is often parameterized in terms of a single parameter related to its variance, hence we call this paradigm variance modeling.

The Gaussian distribution is a popular choice. Assuming that the STFT coefficients of different source components are independent and Gaussian and considering noise as an additional (diffuse) source, the mixture STFT coefficients follow a Gaussian distribution whose covariance matrix $\mathbf{\Sigma}^{\mathbf{X}}_{nf}$ in time-frequency bin (n, f) is given by

$$\mathbf{\Sigma}^{\mathbf{X}}_{nf} = \sum_{j=1}^{J} \mathbf{\Sigma}^{\mathbf{C}}_{jnf} \tag{2.1}$$

where $\mathbf{\Sigma}^{\mathbf{C}}_{jnf}$ is the covariance matrix of the jth source component in that bin. This results in the following expression of the likelihood

$$P\left(\mathbf{X} \mid \mathbf{\Sigma}^{\mathbf{C}}\right) = \prod_{nf} \mathcal{N}\left(\mathbf{X}_{nf}; \mathbf{0}, \sum_{j=1}^{J} \mathbf{\Sigma}^{\mathbf{C}}_{jnf}\right). \tag{2.2}$$

This model is identical to that used by classical beamforming approaches for the enhancement of a target source (Trees, 2002), except that interfering sources are now modeled as individual sources instead of a global background noise. The covariance matrix of each source component can be factored as the product of a scalar nonnegative variance V_{jnf} and a *mixing covariance* matrix $\mathbf{R}_{jnf}$

$$\mathbf{\Sigma}^{\mathbf{C}}_{jnf} = V_{jnf} \mathbf{R}_{jnf} \tag{2.3}$$

which may depend on additional latent variables z via some priors $P(V|z)$, $P(\mathbf{R}|z)$ and $P(z)$. In addition to interchannel intensity and phase differences, mixing covariances encode correlation between channels known as interchannel coherence. As a consequence, the model is not anymore restricted to point sources in low reverberation conditions with high interchannel coherence but becomes valid for diffuse or semi-diffuse sources with lower coherence (El Chami, Pham, Servière, & Guerin, 2008). Under this model, exact MMSE or MAP inference of the source components is often intractable. A common approximation consists of estimating the covariance matrices of the source components in a first stage as

$$\hat{\mathbf{\Sigma}}^{\mathbf{C}}_{\mathrm{MAP}} = \arg\max_{\mathbf{\Sigma}^{\mathbf{C}}} P\left(\mathbf{\Sigma}^{\mathbf{C}} \mid \mathbf{X}\right) \tag{2.4}$$

with

$$P(\mathbf{\Sigma}^{\mathbf{C}} \mid \mathbf{X}) \propto \int P\left(\mathbf{X} \mid \mathbf{\Sigma}^{\mathbf{C}}\right) P\left(\mathbf{R} \mid z\right) P\left(V \mid z\right) P\left(z\right) dz. \tag{2.5}$$

and derive the MMSE STFT coefficients of the source components in a second stage by Wiener filtering

$$\hat{\mathbf{C}}_{jnf} = \hat{\mathbf{\Sigma}}^{\mathbf{C}}_{jnf} \left(\hat{\mathbf{\Sigma}}^{\mathbf{X}}_{nf}\right)^{-1} \mathbf{X}_{nf}. \tag{2.6}$$

This multichannel nonlinear filtering attenuates interfering sounds based both on their spatial direction and their variance (Trees, 2002).

Besides the Gaussian model, another popular model is to assume that the magnitude STFT coefficients of all sources follow a log-Gaussian distribution with tied covariance. Although the likelihood does not admit a closed-form expression, various approximate expressions were proposed in (Roweis, 2001; Kristjánsson, Hershey, Olsen, Rennie, & Gopinath, 2006; Vincent, 2006). In the specific case of single-channel mixtures, Gaussian and Poisson distributions over the magnitude STFT coefficients were also investigated in (Rennie, Hershey, & Olsen, 2008; Virtanen & Cemgil, 2009). Sparse distributions were more rarely considered (Mitianoudis & Davies, 2003; Lee, Kim, & Lee, 2007). In the following, we focus on distributions that can be equivalently defined in terms of a single variance parameter V_{jnf} in the single-channel case and one or more entries of $\mathbf{R}_{jnf}$ in the multichannel case.

Local time-frequency variance models and vector models

Once more, the simplest approach is to assume that the source components have potentially arbitrary spectro-temporal characteristics but distinct spatial distributions. The resulting local time-frequency variance models behave similarly to the local time-frequency linear models presented above. For a point source in low reverberation conditions, the source variances V_{jnf} are equal to the squared magnitude $|S_{jnf}|^2$ of the source signals and the mixing covariances $\mathbf{R}_{jnf}$ are rank-1 matrices equal to the outer product of the mixing vectors $\mathbf{A}_{jnf}$ with themselves

$$\mathbf{R}_{jnf} = \mathbf{A}_{jnf}\mathbf{A}_{jnf}^{H}. \tag{2.7}$$

Some of the priors previously designed for S_{jnf} and $\mathbf{A}_{jnf}$ translate into priors over V_{jnf} and $\mathbf{R}_{jnf}$. For instance, $\mathbf{R}_{jnf}$ was modeled as constant over time and frequency for synthetic instantaneous mixtures in (Févotte & Cardoso, 2005; Vincent, Arberet, & Gribonval, 2009b) and V_{jnf} was modeled by independent binary or sparse discrete priors in (El Chami et al., 2008; Vincent et al., 2009b). Another way of modeling V_{jnf} is to assume that it is constant over small time-frequency regions (Pham, Servière, & Boumaraf, 2003; Févotte & Cardoso, 2005). The parameterization of $\mathbf{R}_{jnf}$ as a rank-1 matrix is not valid for diffuse sources or reverberated point sources, which are better modeled via a full-rank mixing covariance (El Chami et al., 2008). As an extension of the above local time-frequency models, vector models have been proposed that account for across-frequency correlation between the source variances. For instance, V_{jnf} was assumed to be constant over all frequency bins in each time frame in (Mitianoudis & Davies, 2003) and implicit across-frequency variance dependencies were defined using multivariate sparse distributions in (Lee et al., 2007; Lee, 2009).

Approximate MMSE or MAP inference is performed via similar algorithms to those used for local time-frequency linear models. In the case where the number of sources is equal to the number of mixture channels and all sources are point sources, the source components can again be derived from the mixing vectors. ML estimation of the mixing vectors can then be achieved by approximate joint diagonalization of the mixture empirical covariance matrices using a nonstationarity-based FDICA algorithm for local time-frequency models (Pham et al., 2003) or by some *independent vector analysis* (IVA) algorithm for vector models. In the case where the number of sources is larger than the number of mixture channels, MMSE estimation of the source components under a binary variance prior can be addressed via EM and results in soft time-frequency masking (El Chami et al., 2008). Finally, ML estimation of the source covariance matrices with a uniform variance prior is feasible either via EM or nonconvex optimization (Févotte & Cardoso, 2005; Vincent et al., 2009b). Figure 3 illustrates the discrimination potential of variance models with respect to linear models for the separation of a stereo mixture of three sources. As an example, two sets of source components yielding the same mixture STFT coefficients are considered. While the mixture STFT coefficients do not suffice to discriminate these two possible solutions via linear modeling, the additional information carried by the correlation between the mixture channels suffices to discriminate them without any additional prior via variance modeling.

A few systems based on local time-frequency variance models and vector models have been applied to the separation of audio mixtures, focusing on synthetic instantaneous mixtures (Févotte & Cardoso, 2005;

Vincent et al., 2009b) and live recordings of nonmoving sources (Mitianoudis & Davies, 2003; Pham et al., 2003; Lee et al., 2007; El Chami et al., 2008; Lee, 2009). Objective comparisons conducted over stereo instantaneous and anechoic speech and music mixtures in (Puigt et al., 2009; Vincent et al., 2009b) concluded in a SDR improvement compared to nongaussianity-based FDICA and SCA of about 1 dB for mixtures of three sources and more for mixtures of two sources.

Monophonic spectral models

While vector models account for certain variance dependencies across the time-frequency plane, the dependencies exhibited by audio signals are much more complex. Prior distributions commonly used in the field of speech processing rely on the fact that speech is monophonic, *i.e.* it is the result of a single excitation process. The power spectrum V_{jnf} of a given speech source at a given time is assumed to be one of K template spectra w_{jkf} indexed by some discrete state k, which may represent for example a certain phoneme pronounced with a particular intonation. This model extends to music sources which may be considered as a sequence of states representing a certain note or chord played with a particular timbre. The template spectra define the mean power spectrum of each state, which typically involves a set of spectral peaks with characteristic frequencies and intensities. Smooth template spectra parameterizing the coarse spectral structure via an autoregressive (AR) model were also considered in (Srinivasan, Samuelsson, & Kleijn, 2006). The states underlying different sources are modeled as independent and the J-uplet $(q_{1n},\ldots,q_{Jn})$ of states of all sources at a given time is called the factorial state of the mixture. The simplest way of modeling the sequence of states of a single source is to assume that they are independent. Denoting by q_{jn} the state of source j in time frame n and by π_{jk} the prior probability of state k, this yields the mixture model (Attias, 2003; Benaroya, Bimbot, & Gribonval, 2006)

$$V_{jnf} = w_{jq_{jn}f} \text{ with } P(q_{jn} = k) = \pi_{jk}. \tag{2.8}$$

This model is also called *Gaussian mixture model* (GMM) when V_{jnf} is the parameter of a Gaussian or a log-Gaussian distribution. *Hidden Markov models* (HMM) generalize this concept by assuming that each state of a sequence depends on the previous state via a set of transition probabilities ω_{kl} (Roweis, 2001; Kristjánsson et al., 2006)

$$V_{jnf} = w_{jq_{jn}f} \text{ with } \begin{cases} P(q_{j1} = k) = \pi_k \\ P(q_{jn} = l \mid q_{j,n-1} = k) = \omega_{kl} \text{ for } n > 1. \end{cases} \tag{2.9}$$

The transition probabilities may encode both short-term continuity priors and long-term language priors. In the case when the sounds produced by a given source have recurring spectra but variable intensities, its power spectrum may be represented by multiplying each template spectrum by an arbitrary time-varying scale factor h_{jn}. Starting from a GMM, this yields the *Gaussian scaled mixture model* (GSMM) (Benaroya et al., 2006)

$$V_{jnf} = h_{jn} w_{jq_{jn}f} \text{ with } \begin{cases} P(q_{jn} = k) = \pi_{jk} \\ P(h_{jn}) = \mathcal{U}(h_{jn}). \end{cases} \tag{2.10}$$

GMMs, HMMs and GSMMs can apply to multichannel mixtures by specifying a prior over the mixing covariance $\mathbf{R}_{jnf}$ given the source DOA θ_{jn}. A rank-1 prior $P(\mathbf{R}_{jnf} \mid \theta_{jn})$ was employed in (Arberet, Ozerov, Gribonval, & Bimbot, 2009) for instantaneous mixtures and diagonal priors were investigated in (Nix & Hohmann, 2007; Weiss, Mandel, & Ellis, 2008) for live recordings. Priors $P(\theta_j)$ about source movements were also employed in (Nix & Hohmann, 2007).

Different strategies have been used to perform source separation using these models. One strategy employed for single-channel mixtures is to learn via EM the template spectra w_{jkf} and the prior state probabilities π_{jk} or ω_{kl} associated with each source from a set of training signals containing that source alone, then to estimate the MAP scale factors for each possible factorial state on each time frame and either select the optimal factorial state on each frame or find the optimal sequence of factorial states by the Viterbi algorithm (Roweis, 2001; Benaroya et al., 2006; Srinivasan et al., 2006). This necessitates that the identity of the speakers or the musical instruments in the mixture be known and that sufficient training

data is available. Approximate inference techniques based on probability bounds or beam search are used in practice to avoid testing all possible factorial states. Figure 4 illustrates the application of this strategy to a single-channel mixture of piano and violin. The template spectra represent each a single note for violin and a chord consisting of several notes for the piano, where each note translates into a set of peaks at harmonic or near-harmonic frequencies. Although GMMs do not account for frequency or intensity variations within a given note or chord, satisfactory separation is achieved. Another strategy suited to multichannel mixtures is to learn speaker-independent models from a set of training signals covering several speakers and jointly infer the source covariances and the source DOAs via particle filtering or EM (Attias, 2003; Nix & Hohmann, 2007; Weiss et al., 2008). The last and perhaps most promising strategy is to adapt the template spectra and the prior state probabilities to the mixture. Adaptation schemes include MAP selection of one model per source among alternative prior models (Kristjánsson et al., 2006), MAP interpolation between several prior models (Weiss & Ellis, in press) and MAP inference of the model parameters over time intervals involving a single source (Ozerov et al., 2007) or from coarse source estimates provided by a simpler system (Arberet et al., 2009).

GMMs, HMMs and GSMMs have been applied to the separation of single-channel speech mixtures (Roweis, 2001; Kristjánsson et al., 2006; Weiss & Ellis, in press), single-channel music (Benaroya et al., 2006; Ozerov et al., 2007), stereo echoic mixtures of moving or nonmoving speech sources (Attias, 2003; Nix & Hohmann, 2007; Weiss et al., 2008) and stereo instantaneous music mixtures (Arberet et al., 2009). The separation quality achieved over single-channel speech mixtures has not been reported, since most systems focus on estimating the word sequence pronounced by each speaker instead. This typically requires on the order of ten thousand or more template spectra per source and strong language priors, in line with conventional speech recognition systems (Kristjánsson et al., 2006). In (Ozerov et al., 2007), a SDR of 10 dB was achieved on single-channel mixtures of singing voice and accompaniment music using a few tens of states per source only. Multichannel GMMs have been reported to improve SDR by about 2 dB compared to soft time-frequency masking or SCA over stereo echoic mixtures of three speech sources and stereo instantaneous mixtures of three music sources (Weiss et al., 2008; Arberet et al., 2009).

Polyphonic spectral models

Although monophonic spectral models provide a meaningful representation of speech and most musical instruments, they appear less suited to polyphonic musical instruments, which can play chords made of several notes at a time. When the number of concurrent notes increases, the number of states needed to represent all combinations of notes with different relative intensities quickly becomes huge, which favors overfitting. A more principled model consists of representing the power spectrum V_{jnf} of a given source as the linear combination of K basis spectra w_{jkf} multiplied by time-varying scale factors h_{jkn}

$$V_{jnf} = \sum_{k=1}^{K} h_{jkn} w_{jkf} \tag{2.11}$$

where each basis spectrum may represent for example a single note with a certain timbre. The scale factors are often assumed to be independent and follow either uniform priors (Benaroya, McDonagh, Bimbot, & Gribonval, 2003) or sparse priors modeling the fact that only a few basis spectra may be active at the same time (Rennie et al., 2008; Virtanen & Cemgil, 2009). In (Vincent, 2006), dependencies between the scale factors of a given basis spectrum in adjacent time frames were exploited via a binary Markov model. In (Smaragdis, 2007), the basis spectra were replaced by basis spectral patches spanning a few time frames, which is equivalent to setting a constraint over the scale factors of different basis spectra in successive time frames. Priors over the source variances were combined with rank-1 priors over the mixing covariance in (Vincent, 2006; Ozerov & Févotte, 2009).

The Bayesian inference strategies employed in the literature for polyphonic spectral models resemble those for monophonic models. The simplest strategy is again to learn the basis spectra w_{jkf} from separate training data for each source and estimate the scale factors h_{jkn} and additional variables from the mixture (Benaroya et al., 2003; Vincent, 2006; Schmidt & Olsson, 2007; Smaragdis, 2007). For single-channel mixtures, MAP learning and inference consist of approximating an observed matrix of power STFT

coefficients as the product of two nonnegative matrices: a matrix of spectra and a matrix of scale factors. This *nonnegative matrix factorization* (NMF) objective is typically addressed via nonlinear optimization algorithms or via EM. Figure 5 represents the results of this strategy when applied to the single-channel music mixture of Figure 4. The spectrum of the piano is now better modeled as the sum of individual note spectra whose intensities decay over time. However, comparison between the two figures indicates that this increased modeling accuracy does not significantly affect separation performance. Another strategy is to adapt the basis spectra to the mixture by joint MAP inference of all variables. With a single-channel mixture, this requires that priors learned from separate training data be set over the basis spectra of each source (Rennie et al., 2008; Virtanen & Cemgil, 2009). With a multichannel mixture, such priors are not needed provided that the sources have different spatial positions (Ozerov & Févotte, 2009).

Polyphonic spectral models have been applied to the separation of single-channel music mixtures (Benaroya et al., 2003; Vincent, 2006) and multichannel instantaneous and echoic music mixtures of nonmoving sources (Vincent, 2006; Ozerov & Févotte, 2009). In (Vincent, 2006), a SDR of 14 dB was achieved on a highly reverberant stereo mixture of two music sources using knowledge of the instruments and the source DOAs, improving over FDICA and SCA by more than 10 dB. Although polyphonic models are less appropriate for speech than monophonic models, they have also been used for the separation of single-channel speech mixtures since the associated inference algorithms are much faster (Schmidt & Olsson, 2007; Smaragdis, 2007; Rennie et al., 2008; Virtanen & Cemgil, 2009). The resulting SDR over mixtures of two speech sources is typically around 7 dB.

OBJECTIVE PERFORMANCE EVALUATION

While the SDR figures previously reported in this chapter provide an idea of average performance of each class of models, they are not always comparable due to the use of different datasets. Table 1 presents the results achieved by fourteen model-based blind source separation systems on four different types of ten-second speech or music mixtures in the framework of the 2008 Signal Separation Evaluation Campaign (Vincent et al., 2009a). Performance is evaluated not only in terms of SDR, but also in terms of signal-to-interference ratio (SIR) and signal-to-artifacts ratio (SAR), which measure the amount of residual crosstalk and musical noise artifacts.

In low reverberation conditions, the best achieved SDR equals 12 dB on four-channel near-anechoic recordings of three sources and 9 dB on stereo instantaneous mixtures of four sources. The SIR is always larger than the SAR, which indicates satisfactory rejection of interfering sources but linear or nonlinear distortion of the target source. Apart from the lower performance of the system in (Lee et al., 2007) due to convergence issues, the most noticeable difference between systems is that the SDRs achieved by the binary local time-frequency model in (El Chami et al., 2008) and the linear model over a learned signal basis in (Gowreesunker & Tewfik, 2007) are 5 dB lower than those achieved by most other systems. This illustrates the limitations of binary local time-frequency models with respect to continuous models and linear signal bases with respect to spectral variance bases. Interestingly, systems based on monophonic or polyphonic spectral models (Arberet el al., 2009; Ozerov & Févotte, 2009) do not significantly improve performance compared to those based on continuous local time-frequency models (Vincent, 2007). We believe that this may be due to the difficulty of blindly learning spectral models with many parameters from a short mixture without incurring overfitting.

In medium revereberation conditions, the best achieved SDR drops to 4 dB on stereo recordings of two sources and 2 dB on stereo recordings of two sources. The SIR is now always smaller than the SAR, which reveals a lack of discrimination between the sources, and most systems provide almost the same SDR up to 1 dB. Again, the system in (Weiss et al., 2008) based on a spectral model of speech does not improve performance compared to the system in (Mandel et al., 2007) using the same mixing vector priors within a local time-frequency model.

DISCUSSION AND FUTURE RESEARCH DIRECTIONS

General principles and merits of variance modeling vs. linear modeling

The overview of model-based source separation systems conducted in this chapter indicates that most models ultimately follow the same general principle, whereby the mixture STFT coefficients in each time-frequency bin depend on a scalar variable (S_{jnf} or V_{jnf}) encoding spectro-temporal characteristics and a vector or matrix-valued variable ($\mathbf{A}_{jnf}$ or $\mathbf{R}_{jnf}$) encoding spatio-temporal characteristics. This principle defines a link between single-channel and multichannel models. A range of priors have been proposed, relating for example S_{jnf} or V_{jnf} to discrete or continuous latent states and $\mathbf{A}_{jnf}$ or $\mathbf{R}_{jnf}$ to the source DOAs. Intuitively, informative priors over both variables are needed to achieve robust source separation. Indeed, spectral cues alone rarely suffice to discriminate sources with similar pitch range and timbre, while spatial cues alone do not suffice to discriminate sources with the same DOA. Even in situations when either cues suffice to discriminate the sources to a certain extent, combining both cues often increases separation performance (Attias, 2003; Vincent, 2006; Weiss et al., 2008; Arberet et al., 2009).

Two alternative modeling paradigms following the above general principle have been investigated. Linear modeling relies on the assumption that all sources are point sources and appears inappropriate for the encoding of spectral cues. By contrast, variance modeling allows better modeling of reverberant and diffuse sources and efficient encoding of spectral cues. Even when exploiting spatial cues only, variance modeling yields better separation performance on average than linear modeling on mixtures of point sources (Puigt et al., 2009; Vincent et al., 2009b). Therefore we believe that variance modeling is a stronger candidate paradigm for the building of future model-based source separation systems.

Towards higher-level spatial and spectral priors

One of the greatest challenges for future research is to address the separation of difficult mixtures involving a large number of sources, reverberation or source movements. Together with Ellis (2006), we believe that a promising approach is to design stronger priors based on higher-level variables encoding sound features possibly exploited by the auditory system. For instance, mixing covariances could be described via a library of priors relating to distinct mixture acquisition techniques and parameterized as functions of room reverberation time and source directivity in addition to the source DOAs. Priors over the source DOAs themselves could account for typical long-term movements in addition to short-term acceleration. Similarly, the source variances could be represented by separate libraries of priors over the coarse and fine spectral structure providing different parameterizations of periodic and noisy sounds. The variables underlying these priors, including the fundamental frequencies of periodic sounds, would themselves follow different priors for speech and music. Recent advances in this area have been made in the fields of source localization on the one hand (Gustafsson, Rao, & Trivedi, 2003; Fallon, Godsill, & Blake, 2006) and CASA and NMF on the other hand (Ellis, 1996; FitzGerald, Cranitch, & Coyle, 2008).

Towards modular blind source separation systems

Another great challenge for research is to build fully blind systems able to process all kinds of mixtures without any prior information. Many existing systems are designed for a specific number of channels, a specific kind of mixing or a specific kind of sources. Even the systems considered as blind today often rely on prior knowledge of the number of sources, which is rarely available in real-world applications. We believe that modularity and model selection are keys to overcome these limitations. Indeed, a blind system relying on the same spatial and spectral priors over *e.g.* a periodic point source and a noisy diffuse source would likely achieve poor results due to underfitting. Better results might be obtained by providing the system with a library of priors suited to each possible situation and enabling it to select and combine suitable priors depending on the actual situation. The Bayesian framework provides a principled approach to the design of such modular systems (Gelman et al., 2003).

The combination of source models within either modeling paradigm is particularly attractive since the resulting likelihood exhibits a closed-form expression. A step in this direction was taken in (Blouet, Rapaport, Cohen, & Févotte, 2008), where GSMMs, AR models and polyphonic spectral models were combined into a single system achieving MAP inference via a generalized EM algorithm. Experiments conducted on a single-channel mixture of piano and speech showed good performance with manual

selection of a polyphonic spectral model for the piano and an AR model for speech. Additional mixing covariance models and Gaussian variance models, including binary or continuous local time-frequency models, GMMs and HMMs, could also be incorporated into this system using generalized EM inference. Soft masking, FDICA, IVA and spectral model-based Wiener filtering could then all be used at the same time to separate different sources within a given mixture.

In order for such a system to be usable, the most appropriate models should be automatically selected. Advanced Bayesian inference algorithms such as Gibbs sampling and variational Bayes may address this problem by deriving the probabilistic evidence for a particular model (Gelman et al., 2003). These algorithms have already been employed by a few systems for parameter inference within fixed models (Attias, 2003; Cemgil et al., 2007), but not for model selection. Complete evidence maximization could result in a fully blind source separation system, addressing automatic estimation of the number of sources, the most appropriate spatial and spectral model for each source, the number of latent parameters of this model and the best parameter values. Statistical-symbolic languages have recently been proposed that would allow efficient partial implementation of such a system via a sequence of programming constructs (Winn & Bishop, 2005).

CONCLUSION

To sum up, although model-based systems have already achieved very promising separation results, most systems are not entirely blind and fail to separate certain difficult mixtures. The theory of Bayesian signal processing provides a principled research track to address these issues by combination of advanced probabilistic priors and inference criteria. We believe that gradual emergence of the Bayesian modeling paradigms identified in this chapter will result in a major breakthrough in the way source separation systems are developed, with future systems being built from several modules developed by different researchers and selected automatically on the basis of separation quality and computational constraints.

REFERENCES

Arberet, S., Ozerov, A., Gribonval, R., & Bimbot, F. (2009). Blind spectral-GMM estimation for underdetermined instantaneous audio source separation. In *Proceedings of the 8th International Conference on Independent Component Analysis and Signal Separation* (p. 751-758).

Attias, H.. (2003). New EM algorithms for source separation and deconvolution with a microphone array. In *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech and Signal Processing* (p. V-297-300).

Benaroya, L., McDonagh, L., Bimbot, F., & Gribonval, R. (2003). Non negative sparse representation for Wiener based source separation with a single sensor. In *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech and Signal Processing* (p. VI-613-616).

Benaroya, L., Bimbot, F., & Gribonval, R. (2006). Audio source separation with a single sensor. *IEEE Transactions on Audio, Speech and Language Processing, 14(1),* 191-199.

Blouet, R., Rapaport, G., Cohen, I., & Févotte, C. (2008). Evaluation of several strategies for single sensor speech/music separation. In *Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing* (p. 37–40).

Blumensath, T., & Davies, M. E. (2006). Sparse and shift-invariant representations of music. *IEEE Transactions on Audio, Speech and Language Processing, 14 (1),* 50-57.

Cardoso, J.-F. (1998). Multidimensional independent component analysis. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing* (p. IV-1941-1944).

Cardoso, J.-F. (2001). The three easy routes to independent component analysis; contrasts and geometry. In *Proceedings of the 3rd International Conference on Independent Component Analysis and Blind Signal Separation* (p. 1-6).

Cemgil, A. T., Févotte, C., & Godsill, S. J. (2007). Variational and stochastic inference for Bayesian source separation. *Digital Signal Processing, 17 (5),* 891-913.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum-likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39 (1),* 1–38.

El Chami, Z., Pham, D. T., Servière, C., & Guerin, A. (2008). A new model-based underdetermined source separation. In *Proceedings of the 11th International Workshop on Acoustic Echo and Noise Control* (paper ID 9061).

Ellis, D. P. W. (1996). *Prediction-driven computational auditory scene analysis.* Unpublished doctoral dissertation, Dept. of Electrical Engineering and Computer Science, MIT.

Ellis, D. P. W. (2006). Model-based scene analysis. In *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications* (p. 115–146). Wiley/IEEE Press.

Fallon, M. F., Godsill, S. J., & Blake, A. (2006). Joint acoustic source location and orientation estimation using sequential Monte Carlo. In *Proceedings of the 9th International Conference on Digital Audio Effects* (p. 203-208).

Févotte, C., & Cardoso, J.-F. (2005). Maximum likelihood approach for blind audio source separation using time-frequency Gaussian models. In *Proceedings of the 2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (p. 78–81).

FitzGerald, D., Cranitch, M., & Coyle, E. (2008). Extended nonnegative tensor factorisation models for musical sound source separation. *Computational Intelligence and Neuroscience,* article ID 872425.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian Data Analysis, 2nd Edition.* Chapman & Hall/CRC.

Gowreesunker, B. V., & Tewfik, A. H. (2007). Two improved sparse decomposition methods for blind source separation. In *Proceedings of the 7th International Conference on Independent Component Analysis and Signal Separation* (p. 365-372).

Gustafsson, T., Rao, B. D., & Trivedi, M. (2003) Source localization in reverberant environments: modeling and statistical analysis. *IEEE Transactions on Speech and Audio Processing, 11 (6),* 791-803.

Izumi, Y., Ono, N., & Sagayama, S. (2007). Sparseness-based 2ch BSS using the EM algorithm in reverberant environment. In *Proceedings of the 2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (p. 147-150).

Jang, G.-J., & Lee, T.-W. (2004). A maximum likelihood approach to single-channel source separation. *Journal of Machine Learning Research, 4 (7-8),* 1365-1392.

Kristjánsson, T. T., Hershey, J. R., Olsen, P. A., Rennie, S. J., & Gopinath, R. A. (2006). Super-human multi-talker speech recognition: The IBM 2006 speech separation challenge system. In *Proceedings of the 9th International Conference on Spoken Language Processing* (p. 97-100).

Lee, I., Kim, T., & Lee, T.-W. (2007). Independent vector analysis for convolutive blind speech separation. In *Blind speech separation* (p. 169-192). Springer.

Lee, I. (2009). Permutation correction in blind source separation using sliding subband likelihood function. In *Proceedings of the 8th International Conference on Independent Component Analysis and Signal Separation* (p. 767-774).

Makino, S., Lee, T.-W., & Sawada, H. (Eds.). (2007). *Blind Speech Separation.* Springer.

Mandel, M. I., Ellis, D. P. W., & Jebara, T. (2007). An EM algorithm for localizing multiple sound sources in reverberant environments. In *Advances in Neural Information Processing Systems 19* (p. 953-960).

Mitianoudis, N., & Davies, M. E. (2003). Audio source separation of convolutive mixtures. *IEEE Transactions on Speech and Audio Processing, 11 (5),* 489-497.

Mukai, R., Sawada, H., Araki, S., & Makino, S. (2004). Blind source separation for moving speech signals using blockwise ICA and residual crosstalk subtraction. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, E87-A (8),* 1941-1948.

Nesbit, A., Vincent, E., & Plumbley, M. D. (2009). Extension of sparse, adaptive signal decompositions to semi-blind audio source separation. In *Proceedings of the 8th International Conference on Independent Component Analysis and Signal Separation* (p. 605-612).

Nesta, F., Omologo, M., & Svaizer, P. (2008). Separating short signals in highly reverberant environment by a recursive frequency-domain BSS. In *Proceedings of the 2008 IEEE Joint Workshop on Hands-free Speech Communication and Microphone Arrays* (p. 232-235).

Nix, J., & Hohmann, V. (2007). Combined estimation of spectral envelopes and sound source direction of concurrent voices by multidimensional statistical filtering. *IEEE Transactions on Audio, Speech and Language Processing, 15 (3),* 995–1008.

Nocedal, J., & Wright, S. J. (2006). *Numerical Optimization, 2nd Edition.* Springer.

O'Grady, P. D., & Pearlmutter, B. A. (2008). The LOST algorithm: finding lines and separating speech mixtures. *EURASIP Journal on Advances in Signal Processing, 2008,* article ID 784296.

Ozerov, A., Philippe, P., Bimbot, F., & Gribonval, R. (2007). Adaptation of Bayesian models for single-channel source separation and its application to voice/music separation in popular songs. *IEEE Transactions on Audio, Speech and Language Processing, 15 (5),* 1564-1578.

Ozerov, A., & Févotte, C. (2009). Multichannel nonnegative matrix factorization in convolutive mixtures. With application to blind audio source separation. In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing* (p. 3137-3140).

Peterson, J. M., & Kadambe, S. (2003). A probabilistic approach for blind source separation of underdetermined convolutive mixtures. In *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech and Signal Processing* (p. VI-581-584).

Pham, D.-T., Servière, C., & Boumaraf, H. (2003). Blind separation of speech mixtures based on nonstationarity. In *Proceedings of the 7th International Symposium on Signal Processing and its Applications* (p. II–73–76).

Puigt, M., Vincent, E., & Deville, Y. (2009). Validity of the independence assumption for the separation of instantaneous and convolutive mixtures of speech and music sources. In *Proceedings of the 8th International Conference on Independent Component Analysis and Signal Separation* (p. 613-620).

Rennie, S. J., Hershey, J. R., & Olsen, P. A. (2008). Efficient model-based speech separation and denoising using non-negative subspace analysis. In *Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing* (p. 1833-1836).

Roman, N., & Wang, D. L. (2008). Binaural tracking of multiple moving sources. *IEEE Transactions on Audio, Speech and Language Processing, 16 (4),* 728-739.

Roweis, S. T. (2001). One microphone source separation. In *Advances in Neural Information Processing Systems 13* (p. 793-799).

Sawada, H., Araki, S., Mukai, R., & Makino, S. (2007). Grouping separated frequency components with estimating propagation model parameters in frequency-domain blind source separation. *IEEE Transactions on Audio, Speech and Language Processing, 15 (5),* 1592-1604.

Schmidt, M. N., & Olsson, R. K. (2007). Linear regression on sparse features for single-channel speech separation. In *Proceedings of the 2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (p. 26-29).

Smaragdis, P. (1998). Blind separation of convolved mixtures in the frequency domain. *Neurocomputing, 22 (1-3),* 21-34.

Smaragdis, P. (2007). Convolutive speech bases and their application to supervised speech separation. *IEEE Transactions on Audio, Speech and Language Processing, 15 (1),* 1-12.

Srinivasan, S., Samuelsson, J., & Kleijn, W. B. (2006). Codebook driven short-term predictor parameter estimation for speech enhancement. *IEEE Transactions on Audio, Speech and Language Processing, 14 (1),* 163-176.

Trees, H. L. van (2002). *Optimum array processing.* Wiley.

Vincent, E. (2006). Musical source separation using time-frequency source priors. *IEEE Transactions on Audio, Speech and Language Processing, 14 (1),* 91-98.

Vincent, E. (2007). Complex nonconvex lp norm minimization for underdetermined source separation. In *Proceedings of the 7th International Conference on Independent Component Analysis and Signal Separation* (p. 430–437).

Vincent, E., Araki, S., & Bofill, P. (2009a). The 2008 Signal Separation Evaluation Campaign: A community-based approach to large-scale evaluation. In *Proceedings of the 8th International Conference on Independent Component Analysis and Signal Separation* (p. 734-741).

Vincent, E., Arberet, S., & Gribonval, R. (2009b). Underdetermined instantaneous audio source separation via local Gaussian modeling. In *Proceedings of the 8th International Conference on Independent Component Analysis and Signal Separation* (p. 775-782).

Virtanen, T., & Cemgil, A. T. (2009). Mixtures of gamma priors for non-negative matrix factorization based speech separation. In *Proceedings of the 8th International Conference on Independent Component Analysis and Signal Separation* (p. 646-653).

Wang, D. L., & Brown, G. J. (Eds.). (2006). *Computational Auditory Scene Analysis: Principles, Algorithms and Applications.* Wiley/IEEE Press.

Weiss, R. J., Mandel, M. M., & Ellis, D. P. W. (2008). Source separation based on binaural cues and source model constraints. In *Proceedings of the 10th ISCA Interspeech Conference* (pp. 419–422).

Weiss, R. J., & Ellis, D. P. W. (in press). Speech separation using speaker-adapted eigenvoice speech models. *Computer Speech and Language.*

Winn, J., & Bishop, C. M. (2005). Variational message passing. *Journal of Machine Learning Research, 6,* 661–694.

Winter, S., Kellermann, W., Sawada, H., & Makino, S. (2007). MAP-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and l1-norm minimization. *EURASIP Journal on Advances in Signal Processing, 2007,* article ID 24717.

Yılmaz, Ö., & Rickard, S. T. (2004). Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing, 52 (7),* 1830-1847.

Zibulevsky, M., Pearlmutter, B. A., Bofill, P., & Kisilev, P. (2001). Blind source separation by sparse decomposition in a signal dictionary. In *Independent Component Analysis: Principles and Practice* (p. 181-208). Cambridge Press.

APPENDIX: STANDARD PARAMETRIC PROBABILITY DISTRIBUTIONS

$\mathcal{G}$: circular generalized Gaussian distribution over $\mathbb{C}$

$$\mathcal{G}(y;\beta,p) = \frac{p}{2\pi\beta\Gamma(1/p)\,|y|}\exp\left(-\left|\frac{y}{\beta}\right|^{p}\right)$$

with Γ denoting the gamma function.

$\mathcal{N}$: multivariate Gaussian distribution over $\mathbb{C}^I$

$$\mathcal{N}(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\pi^{I/2}\,|\det\boldsymbol{\Sigma}|^{1/2}}\exp\left(-(\mathbf{y}-\boldsymbol{\mu})^{H}\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right)$$

$\mathcal{U}$: uniform distribution

$\mathcal{U}(\mathbf{y})$ is equal to one over the volume of the space over which $\mathbf{y}$ is defined.

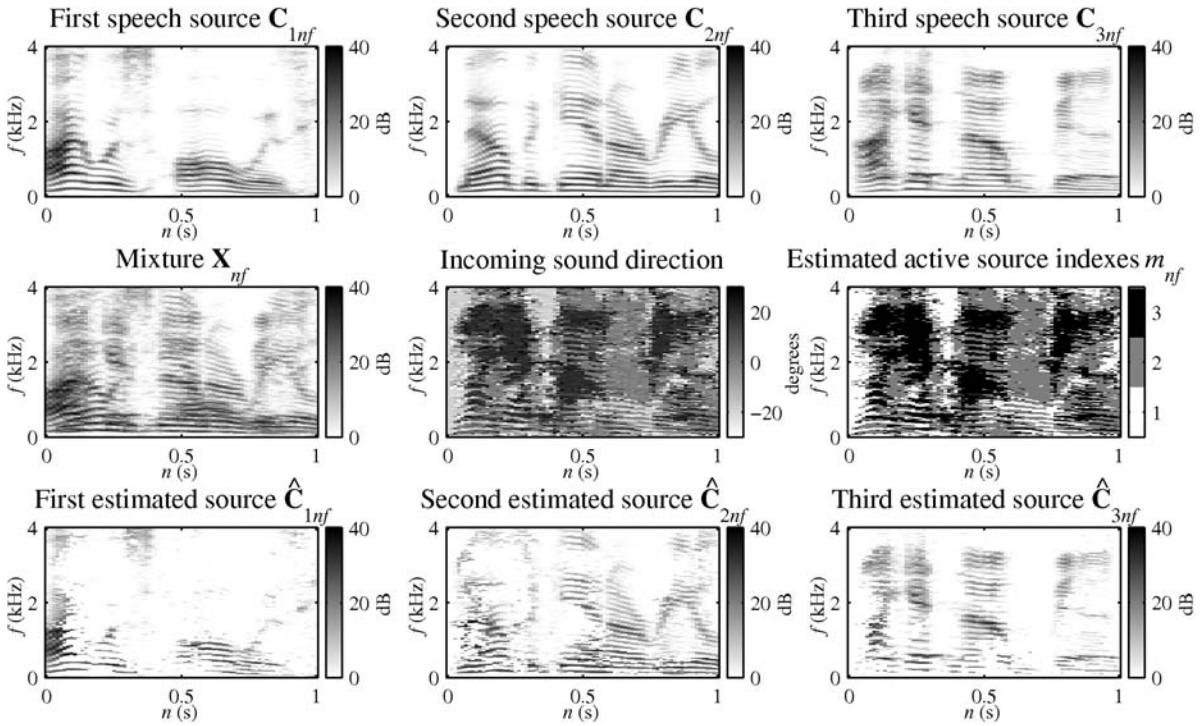

Figure 1. Separation of a stereo instantaneous mixture of three speech sources with DOAs of -20, 0 and 20° using a binary local time-frequency linear model. Only the first channel of each signal is shown.

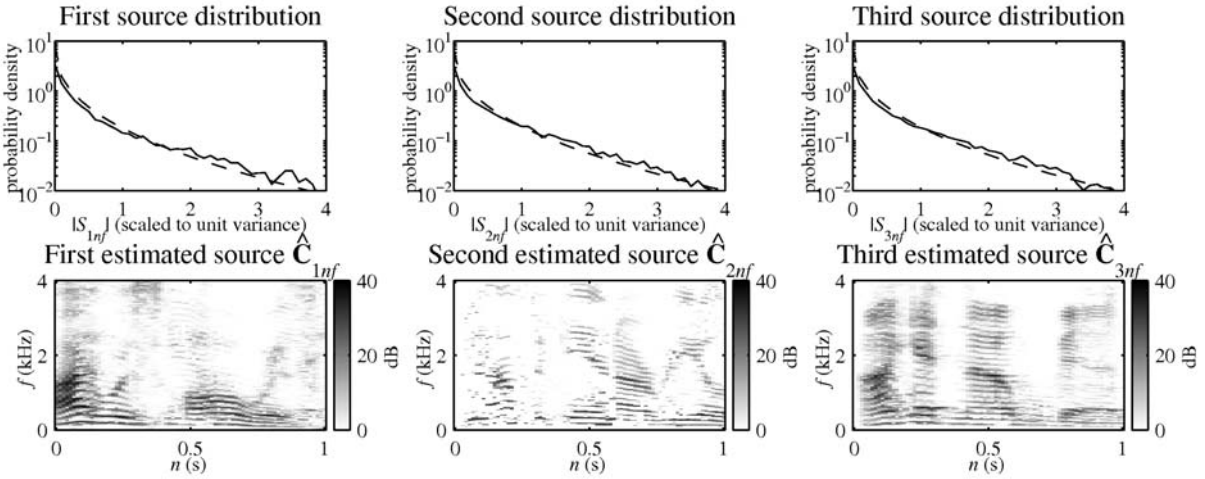

Figure 2. Separation of the stereo instantaneous mixture of Figure 1 using a continuous time-frequency linear model. The distributions of the source magnitude STFT coefficients (plain) are compared to the generalized Gaussian distribution with constant prior scale parameter β and shape parameter p=0.4 (dashed). Only the first channel of each estimated source component is shown.

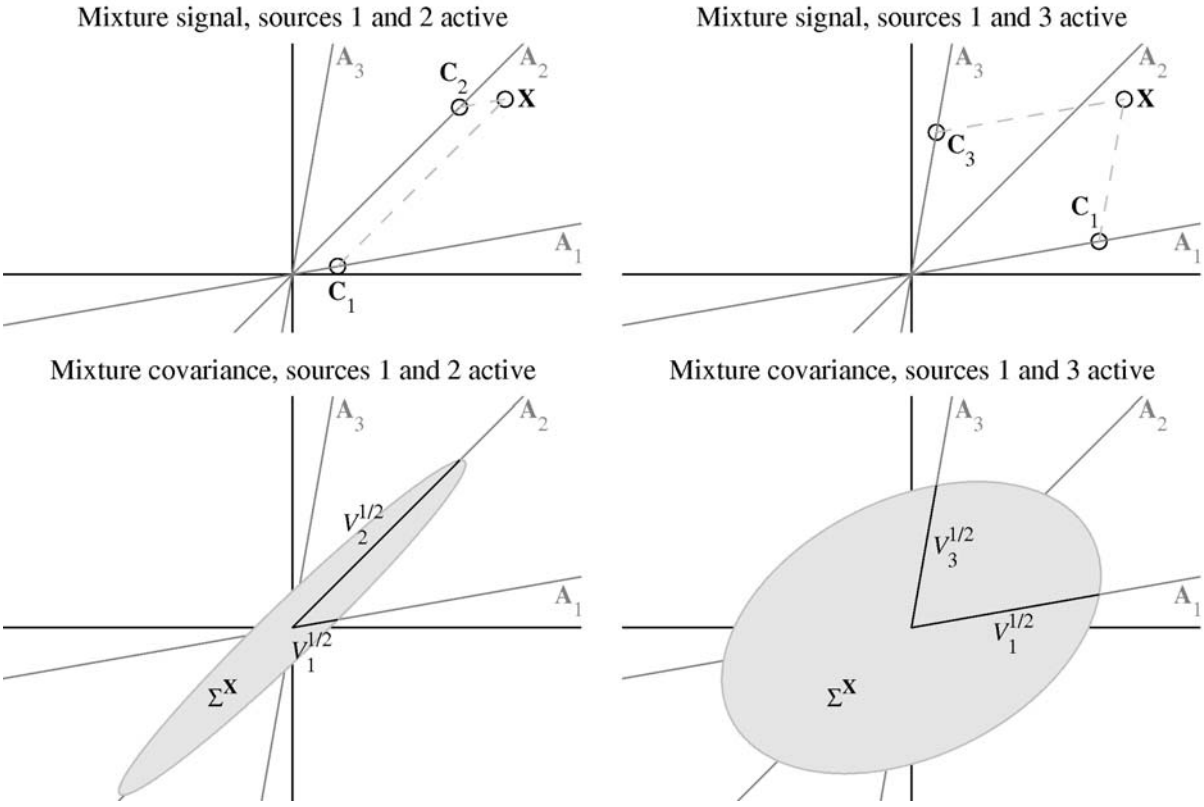

Figure 3. Illustration of the discrimination between two possible solutions (left and right) to the separation of a stereo mixture of three sources via local time-frequency linear modeling (top) vs. local time-frequency variance modeling (bottom) in a given time-frequency bin. Covariance matrices are represented as ellipses whose axes and axe lengths represent their eigenvectors and the square roots of their eigenvalues. All variables are represented as real-valued and time-frequency indices are dropped for the sake of legibility.

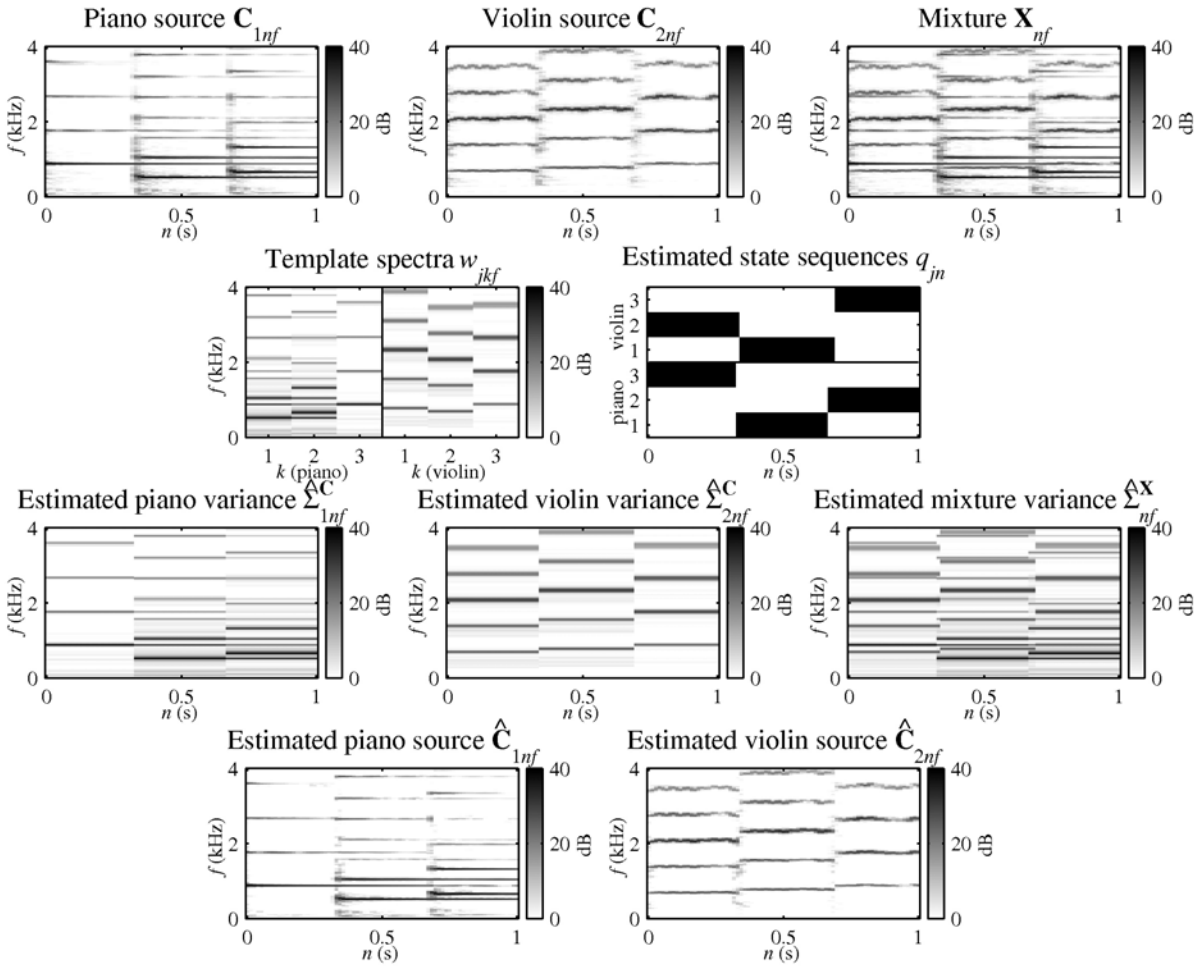

Figure 4. Separation of a single-channel mixture of two music sources using GMMs.

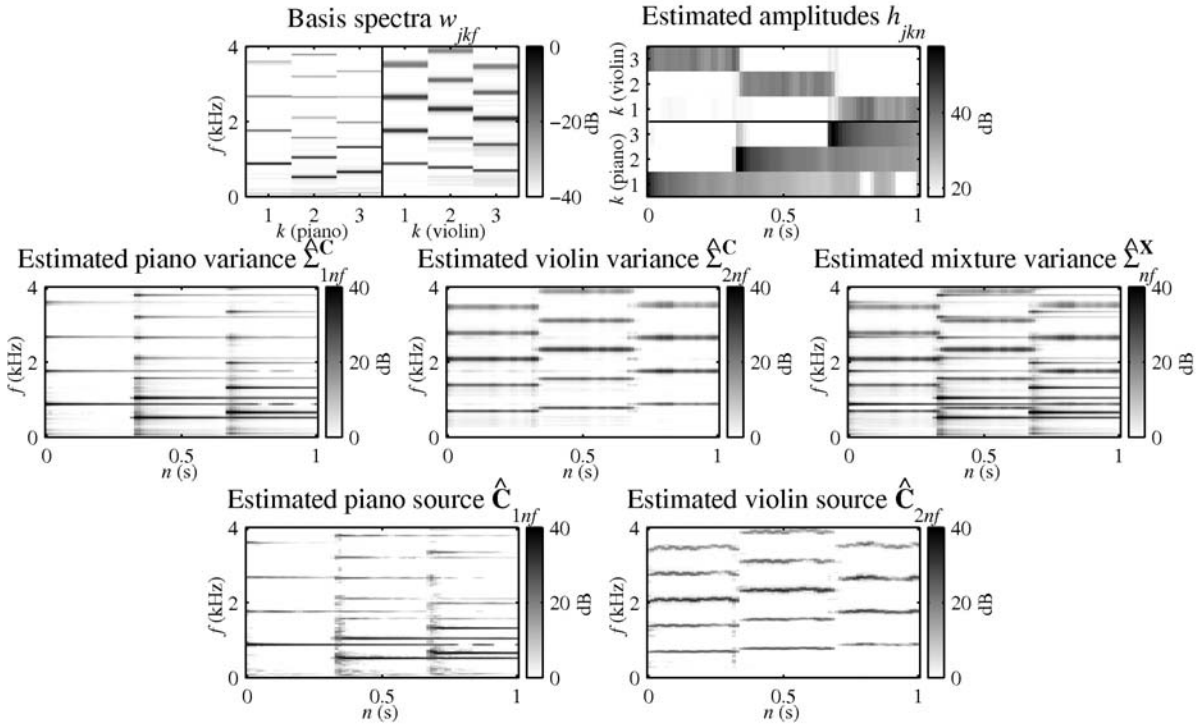

Figure 5. Separation of the single-channel mixture of two music sources in Figure 4 using a polyphonic spectral model.

System	Average separation quality		
	SDR (dB)	SIR (dB)	SAR (dB)
Four-channel recordings of three music sources in a room with cushioned walls			
(Lee et al., 2007)	0.7	6.1	4.1
(Lee, 2009)	9.8	18.8	10.5
(Nesta et al., 2008)	11.7	23.0	12.3
Binaural recordings of two speech sources in an office room			
(Mandel et al., 2007)	4.4	5.9	11.2
(Weiss et al., 2008)	3.7	6.1	10.7
Stereo recordings of four speech sources in an office room			
(El Chami et al., 2008)	1.9	1.4	7.4
(Izumi et al., 2007)	1.6	1.4	6.6
(Mandel et al., 2007)	0.9	-3.3	12.5
(Sawada et al., 2007)	2.3	4.2	5.0
Stereo instantaneous mixtures of four speech sources			
(Arberet et al., 2009)	8.1	14.5	8.8
(El Chami et al., 2008)	3.9	10.8	7.4
(Gowreesunker & Tewfik, 2007)	3.6	12.6	4.8
(Nesbit et al., 2009)	6.4	11.5	7.3
(Ozerov & Févotte, 2009)	8.6	14.8	9.7
(Vincent, 2007)	8.3	14.6	9.1
(Vincent et al., 2009b)	8.0	13.8	8.9

Table 1. Average source separation performance achieved by model-based systems in the framework of the 2008 Signal Separation Evaluation Campaign.

[i] In the particular case when all the sources are point sources, the source separation problem is often defined as that of recovering the single-channel source signals emitted by individual sources. Strictly speaking, this involves not only separation, but also dereverberation. Moreover, these source signals can be estimated at best up to a filtering indeterminacy in the absence of information about *e.g.* preamplifier transfer functions or global sound effects. The chosen definition of the source separation problem in terms of source components, taken from (Cardoso, 1998), is more general, since it is valid for diffuse sources and it is not subject to the above indeterminacy.