Model-Based Audio Source Separation

EMMANUEL VINCENT, MARIA G. JAFARI, SAMER A. ABDALLAH,
MARK D. PLUMBLEY AND MIKE E. DAVIES

Technical Report C4DM-TR-05-01
Revised version, 25 october 2006

Model-Based Audio Source Separation

Revised version, 25 october 2006
Former title: Blind Audio Source Separation

Emmanuel Vincent, Maria G. Jafari, Samer A. Abdallah, Mark D. Plumbley

Centre for Digital Music
Queen Mary, University of London
Mile End Road – London E1 4NS – United Kingdom
emmanuel.vincent@elec.qmul.ac.uk

Mike E. Davies

Institute for Digital Communication
University of Edinburgh
Mayfield Road – Edinburgh EH9 3JL – United Kingdom
mike.davies@ed.ac.uk

Abstract: Most audio signals are mixtures of several audio sources which are active simultaneously. Audio source separation is the problem of recovering each source signal from a given mixture signal. Historically, audio source separation systems relied on beamforming algorithms, which do not require any prior knowledge about the source signals and can be applied whenever the mixture is recorded from a set of microphones with known relative positions. Their performance is often very good when the number of microphones is large, but it decreases quickly when the number of microphones is small. An alternative approach is to rely on models of the source signals to make better use of the available information. Existing models rely on some form of independence of the sources along with other assumptions. This report provides a tutorial review of model-based audio source separation algorithms, focusing on situations where the number of mixture channels is limited and possibly smaller than the number of sources. We highlight the exact assumptions made by each algorithm and discuss their validity and limitations for real-world audio signals. To this aim, approaches relating to different historical viewpoints are interpreted within a general statistical framework. We do not discuss implementation issues, but provide bibliographical and software references for more details.

Keywords: source separation, music, speech, recording, CD, multichannel filtering, time-frequency masking, independent component analysis, DUET, factorial hidden Markov model, nonnegative matrix factorization, computational auditory scene analysis.

Most audio signals are mixtures of several audio 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 each source signal from a given mixture signal.

This problem is relevant to many applications where the separated sources are directly listened to, including speech enhancement for hearing aids and sampling of instrumental sounds for electronic music composition. Moreover, the source signals may be remixed differently to create a new mixture signal for a specific application such as voice cancellation for karaoke, rendering of stereo CDs on multichannel devices or post-production of raw music recordings. The estimated source signals may also be fed into single-source indexing, transcription and coding techniques, allowing improved music indexing, multi-speaker speech recognition or object-based coding.

Historically, audio source separation systems relied on *beamforming* algorithms such as the delay-and-sum (DS) beamformer, which enhances sounds from a target direction regardless of the spatial distribution of interference, and the generalized sidelobe canceller (GSC), designed to minimize the energy of interference without distorting sounds from the target direction. These algorithms have been the subject of many reviews, *e.g.* [1, 2]. They do not require any prior knowledge about the source signals and can be applied whenever the mixture is recorded from a set of microphones with known relative positions. Their performance is often very good when the number of microphones is large. However, it decreases quickly in the presence of reverberation or many interference sources when the number of microphones is small [3, 4].

Since the recording setup cannot always be chosen in practice, alternative source separation methods are needed when few microphones are available or when the sources are mixed synthetically with a mixing desk. An interesting approach is to rely on *models* of the source signals to make better use of the available information. This is consistent with listening experiments [5] proving that the ability of the human auditory system to segregate sound sources based on their spatial direction degrades with reverberation, and that additional cues such as harmonicity and temporal structure become increasingly important in this context. Early model-based approaches were proposed independently about twenty years ago by researchers in neural networks [6] and computational perception [7] and have led to a wide range of source separation methods today. These methods generally rely on some form of *independence* of the sources along with other assumptions. Methods involving generic models which apply to all types of sources are often termed *blind* source separation methods. By contrast, methods exploiting prior information about the sources in a particular mixture are called *semi-blind*, although the distinction between blind and semi-blind methods is sometimes unclear. Typically, a wide range of models can be used to separate a given mixture.

In this article, we provide a tutorial review of model-based audio source separation algorithms, focusing on situations where the number of mixture channels is limited and possibly smaller than the number of sources. We highlight the exact assumptions made by each algorithm and discuss their validity and limitations for real-world audio signals. To this aim, approaches relating to different historical viewpoints are interpreted within a general statistical framework. We do not discuss implementation issues, but provide bibliographical and software references for more details. Also we do not consider other problems related to audio source separation such as dereverberation and remixing.

The rest of the article is structured as follows. In the first section, we describe the properties of audio mixtures exploited within signal models, state a formal definition of the audio source separation problem and show that it is equivalent to the identification of the parameters of a chosen separating function. We present some popular families of separating functions in the second section. In the three subsequent sections, we provide a review of established and recent algorithms used to identify the parameters of these functions, sorted approximately in order of increasing complexity of the underlying models. In the final section, we summarize the similarities between these algorithms, their performance and their relation to beamforming and we conclude by discussing future challenges.

1 Background and definitions

1.1 Audio sources

Audio sources can be categorized as speech, music or environmental sounds. Each category exhibits specific spectro-temporal characteristics that can be exploited for source separation.

Speech [8] can be analyzed as a sequence of discrete units called *phonemes*. The signal corresponding to each phoneme exhibits nonstationary characteristics and can include a periodic part containing harmonic sinusoidal partials,

a noise part or a transient part. A few phonemes contain a superposition of periodic and noisy components. The harmonicity property means that the frequencies of the sinusoidal partials are multiples of a single frequency called the *fundamental frequency*. The fundamental frequency varies over time, but stays within a range of about 40 Hz centered around an average of 140 Hz for male and 200 Hz for female speakers. The *spectral envelope*, that is the smoothed profile of the magnitude spectrum, depends on the phoneme, the phonetic context and the speaker identity. Successive phonemes build up into words and sentences governed by language-specific rules, with silence intervals separating sentences. Interfering speech sources are usually unrelated to the target source.

Music sources [9], which include musical instruments and singers, produce sequences or superpositions of events termed *notes* or tones. The signal corresponding to each note is composed of a transient part, possibly followed by a near-periodic part containing harmonic sinusoidal partials. Some instruments produce additional noise. In western music, the fundamental frequency of a note is generally constant or slowly varying around a value on the *semitone* scale, that is a discrete logarithmic scale spanning the range between 30 Hz and 4 kHz. Each instrument is characterized by a specific *timbre* which depends mostly on the duration of onset transients, the spectral envelope and the amount of frequency modulation. Successive notes are often played without any silence in between. Within a music ensemble, western harmony rules favor synchronous notes at rational fundamental frequency ratios such as 2, $3/2$ or $5/4$. Thus harmonic partials from different sources often overlap at some frequencies.

Environmental sounds [10] have different characteristics depending on their origin. They may involve periodic, noisy or transient signals and follow a simpler temporal organization than speech or music.

1.2 Live recordings and synthetic mixtures

Mixtures of audio sources can be acquired either by live recording or by synthetic mixing, as illustrated in Figure 1. Since synthetic mixing effects differ from the physics of recording, the resulting mixture signals have noticeably different spatial properties that are exploited by some source separation methods. The observed spatial properties also depend on the spatial extent of the sources. Speakers and small musical instruments can be modeled as *point sources* producing sound from a single point in space. Larger instruments such as piano or drums, which can produce sound at different spatial positions at the same time, are called *extended sources*.

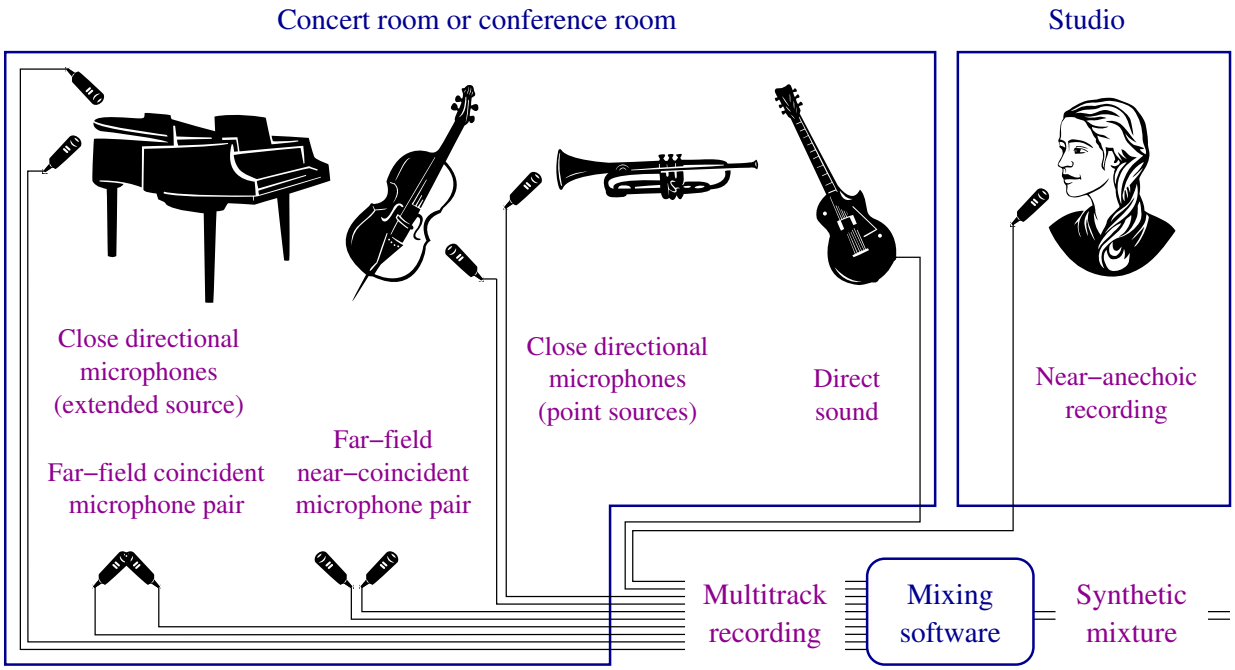

Figure 1: The two ways of obtaining audio mixtures: live recording (left) and synthetic mixing (bottom right).

"Cocktail party" speech mixtures are generally obtained by recording all the sources simultaneously using multiple microphones. Common microphone arrangements may involve both near-field and far-field microphones with various directivities [11]. The recorded signals contain filtered versions of the sources due to successive reflections of the

sound waves on the room surfaces [12]. *Binaural* recordings obtained from one microphone in each ear are also often employed for hearing aid applications and involve additional filtering of the sources due to the shape of the ears and the head. The positions of the first few early reflections may be predicted to some extent from the geometry of the room. Late reflections, known as *reverberation*, appear at diffuse near-random locations. The amount of reverberation can be measured by the reverberation time RT_{60}, which is the delay after which the magnitude of the reflections becomes 60 dB smaller than that of the original sound. This quantity is on the order of 150 to 500 ms in office rooms and 1 to 2 s in concert halls. The length of the room impulse response between a point source and a microphone is proportional to the reverberation time and smaller for near-field microphones than far-field ones. Room impulse responses generally vary over time due to source movements, and even small uncontrolled movements may result in large variations of reverberation.

Pop music CDs and movie soundtracks are often made not from such live recordings, but instead by recording the sources separately in a near-anechoic studio with a single microphone, then applying different special effects to each source and mixing them together using a mixing desk or dedicated software [11]. The mixture signal is typically in *stereo* (two-channel) or some other multichannel format. The mixing process consists of transforming each single-channel source signal into a multichannel *spatial image* with a chosen, possibly moving, spatial position using one or more effects such as "pan", which scales the signal by channel-specific positive gains, and "reverb", which simulates reverberation by applying channel-specific synthetic filters. These spatial images are then added together on each channel to obtain the mixture signal. For more information about the many effects available, see the documentation of mixing software such as TAP-plugins[1]. Synthetic mixing may also be used to merge the channels of a live recording. This technique employed for classical music CDs results in very complex spatial properties due to the combination of both natural and synthetic mixing effects.

1.3 Formalization of the mixing process and categories of mixtures

The mixing process for live recordings and synthetic mixtures may be formalized in the same way. In the following, we assume that all the signals are sampled at a fixed frequency. We define I and J as the number of mixture channels and sources respectively and we use square brackets to denote vectors and matrices. When only point sources are present, the channels $[x_i(t)]_{1\leq i\leq I}$ of the mixture signal can be expressed by $x_i(t) = \sum_{j=1}^{J} \sum_{\tau=-\infty}^{+\infty} a_{ij}(t-\tau,\tau) s_j(t-\tau)$ in the time domain where $[s_j(t)]_{1\leq j\leq J}$ are the single-channel source signals and $[a_{ij}(t,\tau)]_{1\leq i\leq I, 1\leq j\leq J}$ is a set of time-varying *mixing filters* describing either the room impulse responses or the synthetic mixing effects or both. This expression does not hold for extended sources, which cannot be represented as single-channel signals. More generally, the mixing process can always be written as [13]

$$x_i(t) = \sum_{j=1}^{J} s_{\mathrm{img}\,ij}(t) \tag{1}$$

where $s_{\mathrm{img}\,ij}(t)$ is the spatial image of source j on mixture channel i. This quantity is defined for a point source by

$$s_{\mathrm{img}\,ij}(t) = \sum_{\tau=-\infty}^{+\infty} a_{ij}(t-\tau,\tau) s_j(t-\tau) \tag{2}$$

and it can also be defined for an extended source by integrating (2) over the source spatial domain. These equations may be mapped to the time-frequency domain using the complex short-term Fourier transform (STFT). Denoting by $X_i(n,f)$, $S_j(n,f)$ and $S_{\mathrm{img}\,ij}(n,f)$ the STFT coefficients of $x_i(t)$, $s_j(t)$ and $s_{\mathrm{img}\,ij}(t)$ respectively in time frame n and frequency bin f and approximating the mixing filters by complex mixing matrices $[A_{ij}(n,f)]_{1\leq i\leq I, 1\leq j\leq J}$ under the *narrowband assumption* [1], this gives

$$X_i(n,f) = \sum_{j=1}^{J} S_{\mathrm{img}\,ij}(n,f) \tag{3}$$

$$S_{\mathrm{img}\,ij}(n,f) \approx A_{ij}(n,f) S_j(n,f). \tag{4}$$

[1] `http://tap-plugins.sourceforge.net/ladspa.html`

In the context of source separation, mixtures are often broadly categorized according to three criteria related to the separation difficulty: the relative number of mixture channels and sources, the length of the mixing filters and the time variation of the mixing filters [14]. A mixture is termed *over-determined* when the number of channels I is larger than the number of sources J, *determined* when it is equal and *under-determined* when it is smaller. It is also termed *instantaneous* when the mixing filters are simple zero-delay gains a_{ij}, *anechoic* when they represent additional (possibly fractional) delays τ_{ij}, and *convolutive* otherwise. Finally, a mixture with static sources or fixed mixing filters $a_{ij}(\tau)$ is called a *time-invariant* mixture, while one with moving sources or time-varying filters $a_{ij}(t, \tau)$ is called *time-varying*. These terms will be employed throughout the rest of this article.

1.4 Definition of the source separation problem

The source separation problem consists of estimating the source spatial images $[s_{\mathrm{img}\,ij}(t)]_{1\leq i\leq I, 1\leq j\leq J}$ from the mixture signal $[x_i(t)]_{1\leq i\leq I}$. In the absence of source-specific prior information, this is feasible at best up to an arbitrary ordering. When the mixture involves point sources, separation may be followed by dereverberation, that is the estimation of the underlying single-channel source signals $[s_j(t)]_{1\leq j\leq J}$. This additional problem, which is not part of the source separation problem strictly speaking, is not considered in the following. Most source separation methods can estimate single-channel source signals only up to a gain or filtering indeterminacy, which disappears when considering source spatial images instead [13, 15].

In practice, perfect separation is rarely achieved and the estimated source spatial images may contain different kinds of distortion including *interference* from other sources, *musical noise*, timbre distortion and spatial distortion. Musical noise, which consists of "gurgling" artifacts appearing with some time-varying filtering algorithms, is often considered more annoying than other distortions when the estimated signals are destined to be listened to [16]. Currently, these distortions may be quantified precisely only by means of listening tests. Objective criteria expressed in decibels based on the knowledge of the true mixing filters or the true source spatial images are often used instead for convenience [17, 18]. In the following, we report the signal-to-noise ratio (SNR) or the signal-to-interference ratio (SIR) when available.

Mathematically, the separation process consists of finding a data-adaptive function that takes a mixture signal $[x_i(t)]_{1\leq i\leq I}$ as input and outputs source spatial image estimates $[\widehat{s}_{\mathrm{img}\,ij}(t)]_{1\leq i\leq I, 1\leq j\leq J}$. Acceptable functions are usually restricted to parametric functions of the form $[\widehat{s}_{\mathrm{img}\,ij}(t)]_{1\leq i\leq I, 1\leq j\leq J} = g([x_k(t)]_{1\leq k\leq I}, \theta)$, where g is some fixed separating function and θ a vector of parameters. Hence the separation problem can be broken into two steps: firstly choice of a suitable separating function g, and secondly adaptation of the parameters θ to the observed mixture signal according to some model. These two steps are now discussed successively.

2 Parametric separating functions

Classically, two assumptions can be made in order to separate a mixture signal: *spatial diversity* and/or *time-frequency diversity*. Spatial diversity means that the sources are located in different regions of space, whereas time-frequency diversity means that the sources are active in different regions of the time-frequency plane. Three families of parametric separating functions can be derived from these assumptions: multichannel time-invariant filtering, channel-wise time-frequency masking and multichannel time-varying filtering.

2.1 Multichannel time-invariant filtering

Multichannel time-invariant filtering is the process of convolving the mixture channels by a set of time-invariant filters called *demixing filters* and summing the filtered channels together. This process exploits spatial diversity by acting as a spatial filter that attenuates sounds from certain directions of arrival depending on frequency [1]. It is most often expressed in the time-frequency domain via the STFT. The demixing filters are then defined by a matrix of complex coefficients $[W_{ji}(f)]_{1\leq i\leq I, 1\leq j\leq J}$ for each frequency bin f which must be adapted to the observed mixture. The STFT $S_j(n, f)$ of the single-channel source signal $s_j(t)$ is estimated from the STFTs $[X_i(n, f)]_{1\leq i\leq I}$ of the mixture

channels $[x_i(t)]_{1 \leq i \leq I}$ by

$$\widehat{S}_j(n,f) = \sum_{i=1}^{I} W_{ji}(f) X_i(n,f) \tag{5}$$

and the STFT $S_{\text{img}\,ij}(n,f)$ of the source spatial image $s_{\text{img}\,ij}(t)$ by

$$\widehat{S}_{\text{img}\,ij}(n,f) = B_{ij}(f) \widehat{S}_j(n,f) \tag{6}$$

where $[B_{ij}(f)]_{1 \leq i \leq I, 1 \leq j \leq J}$ is the pseudo-inverse of $[W_{ji}(f)]_{1 \leq i \leq I, 1 \leq j \leq J}$ [15]. The corresponding time-domain signal $\widehat{s}_{\text{img}\,ij}(t)$ is derived by STFT inversion. The estimated source spatial images remain well-defined even when the demixing coefficients are specified up to an arbitrary complex multiplicative coefficient for each source j and each frequency bin f [13, 15]. Circular convolution effects may be avoided by zero-padding of the FFT and corresponding constraints on the coefficients [19].

The main advantage of time-invariant filtering is that it does not generate musical noise [2]. In theory, determined or over-determined time-invariant mixtures involving point sources only can be perfectly separated by constructing the demixing filter system to be the inverse of the mixing filter system. When the mixing filters have a finite length, the optimal demixing filters also have a finite length in the over-determined case, but may be infinite otherwise [20]. The optimal demixing filters are often much longer than the mixing filters, as illustrated in Figure 2. Shorter filters of the order of a thousand taps may still achieve good separation for determined time-invariant mixtures generated by mixing filters of similar length [21]. The performance of multichannel time-invariant filtering decreases on under-determined mixtures or reverberant mixtures because the number of directions of arrival that can be perfectly cancelled is limited by the number of mixture channels, so that only partial interference attenuation becomes feasible [4, 21]. Performance also decreases on time-varying mixtures or mixtures involving extended sources for the same reason [22].

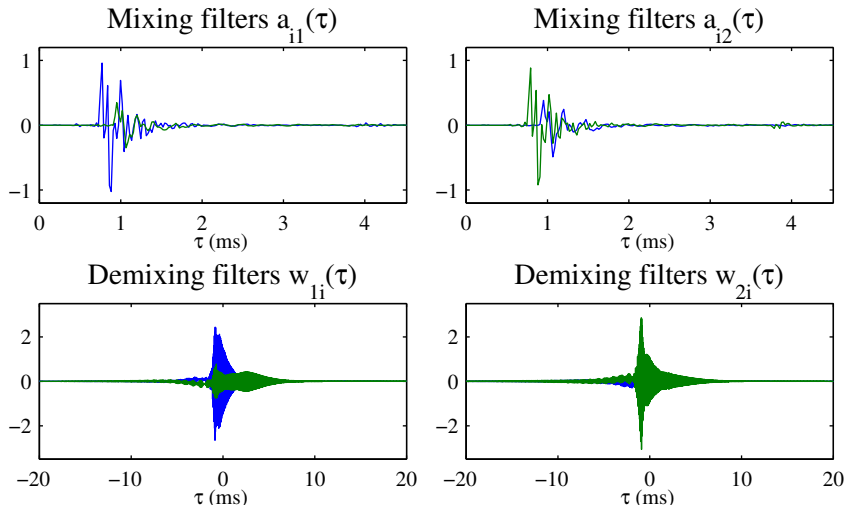

Figure 2: Binaural mixing filters for two static sources at -20 and +20° in a near-anechoic environment and optimal demixing filters computed by inversion of the mixing filter system and truncated between -20 and +20 ms.

2.2 Channel-wise time-frequency masking

Channel-wise time-frequency masking is a special case of time-varying filtering conducted on each mixture channel separately. This technique relies on time-frequency diversity and can be employed both for single-channel and multichannel mixtures to attenuate sounds in certain regions of the time-frequency plane [23, 24]. It is most often expressed using a STFT representation, although other time-frequency-like representations such as auditory filterbanks can also be employed [25, 26]. The STFT $S_{\text{img}\,ij}(n,f)$ of the source spatial image $s_{\text{img}\,ij}(t)$ is estimated by

$$\widehat{S}_{\text{img}\,ij}(n,f) = M_j(n,f) X_i(n,f) \tag{7}$$

where $M_j(n,f)$ is a *time-frequency mask* containing positive gains which must be adapted to the observed mixture. The corresponding time-domain signal $\widehat{s}_{\text{img}\,ij}(t)$ is derived by STFT inversion. The set of possible gains is sometimes restricted to $M_j(n,f) \in \{0,1\}$, leading to the so-called *binary masking* technique [23, 24].

In theory, binary masking can provide perfect source separation even with under-determined, reverberant or time-varying mixtures provided the source spatial images are *disjoint orthogonal* [24], *i.e.* at most one source is active in any time-frequency bin. In practice, this is only partially true and musical noise may be heard where the sources overlap in time-frequency [16]. The optimal STFT length which maximizes disjoint orthogonality has been found to be 60 ms for speech [27] and 200 ms for music [28] on average. Optimal binary masks associating each time-frequency bin with the dominant source have been reported to provide good performance on under-determined anechoic speech mixtures [27] and to improve speech intelligibility, despite the introduction of musical noise [26]. Non-binary masks may also produce musical noise, but it can be reduced using temporal smoothness [16], spectral smoothness [29] or psycho-acoustical [30] constraints on the masks. Example non-binary masks are plotted in Figure 3.

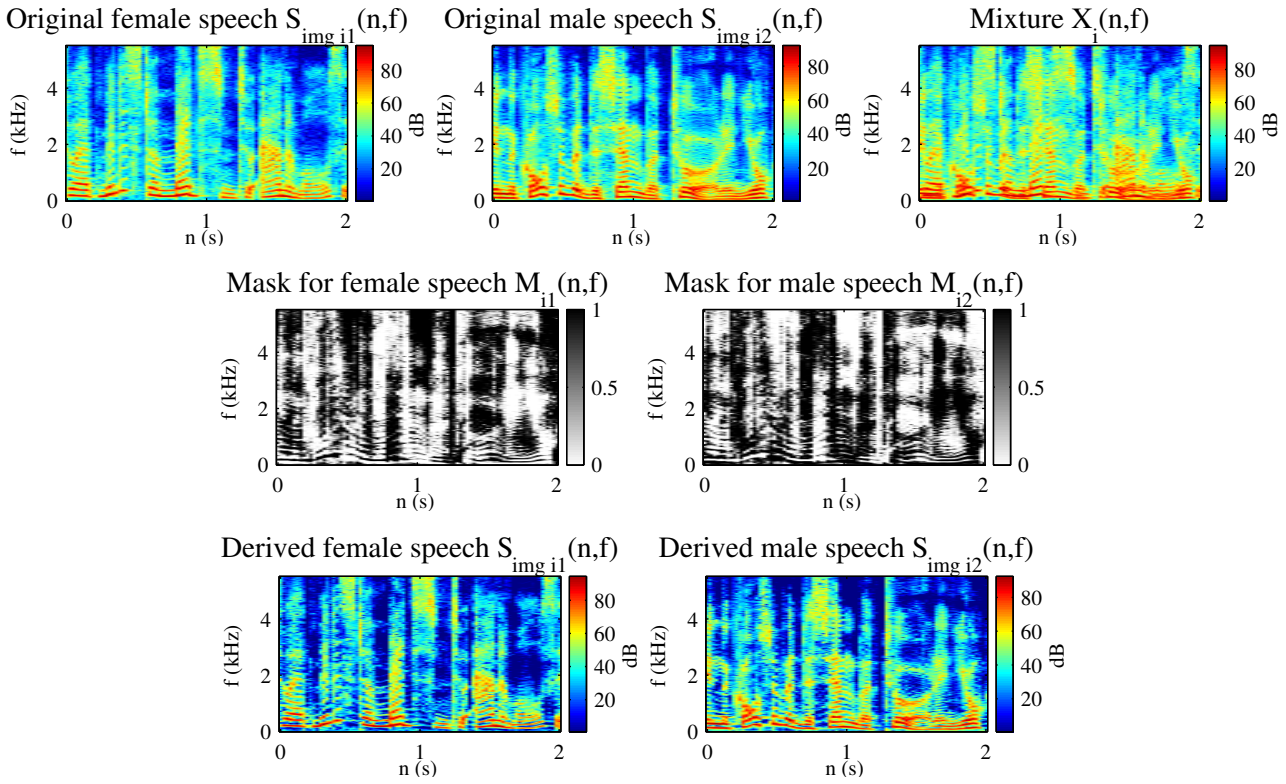

Figure 3: Optimal non-binary time-frequency masks for a speech mixture computed according to the Wiener filtering formula [16] and derived source estimates. Dark parts of the masks indicate selected time-frequency regions.

2.3 Multichannel time-varying filtering

Multichannel time-varying filtering is a generalization of multichannel time-invariant filtering and channel-wise time-frequency masking where the mixture channels are convolved by time-varying demixing filters and added together. This allows the joint exploitation of spatial diversity and time-frequency diversity. In the STFT domain, the filtering process is defined by a matrix of complex demixing coefficients $[W_{ji}(n,f)]_{1\leq i\leq I, 1\leq j\leq J}$ for each time-frequency bin (n,f), which must be adapted to the observed mixture. The source spatial images are then estimated using equations (5) and (6) with an additional dependence on the time frame n in $W_{ji}(n,f)$ and $B_{ij}(n,f)$.

Due to the large number of demixing coefficients, additional constraints are often assumed to make the adaptation of these coefficients more tractable. The *generalized disjoint orthogonality* constraint [31, 32] states that at most N sources are active in each time-frequency bin, where N is typically set either to one [33, 27], the number of mixture channels I [34] or any number smaller than I [32]. The adaptation of the demixing coefficients then consists of

estimating the set of active sources and the associated demixing coefficients, which are assumed to be real-valued for instantaneous mixtures [31, 34]. Increasing N can result in a large improvement of the separation performance for under-determined mixtures, both in the instantaneous case and in the convolutive case with large STFT length [28]. Another common constraint is that the demixing coefficients are constant over blocks of time frames [35].

3 Multichannel separation based on spatial cues

Once a parametric separating function has been chosen, the source separation problem becomes that of adapting the demixing filters or the time-frequency masks to the observed mixture. The simplest adaptation methods segregate the sources in multi-channel mixtures relying on spatial cues. Their core modeling assumptions are that the source signals are statistically independent and sparse. Here *sparsity* refers to the property by which most of the sample values are close to zero. This is true for many speech signals in the time domain and both for speech and music signals in the time-frequency domain [34, 36, 27]. An additional assumption that the source magnitudes are correlated across frequency may be used to separate convolutive mixtures [15]. Two families of algorithms are reviewed in this section: convolutive independent component analysis and directional masking.

3.1 Convolutive independent component analysis

Independent component analysis (ICA) and its extensions aim to separate determined or over-determined mixtures by multichannel filtering, based on the main assumption that the source signals are statistically independent. Basic ICA algorithms focus on determined time-invariant instantaneous mixtures, where the mixing filters reduce to a zero-delay square mixing matrix $[a_{ij}]_{1\leq i\leq I,1\leq j\leq I}$ that can be inverted in the time domain using a demixing matrix $[w_{ji}]_{1\leq i\leq I,1\leq j\leq I}$. These algorithms have been extensively reviewed in *e.g.* [37, 38]. The optimal demixing matrix is identifiable up to arbitrary row permutation and scaling provided that at most one source is a stationary white Gaussian noise and the mixing matrix is invertible, which implies that the sources have different spatial positions. It can be estimated using probabilistic source models by minimizing the mutual information between the resulting source signals or by maximizing their probability given the mixture signal [37], using gradient ascent or other optimization techniques. The most popular model assumes that the samples $s_j(t)$ of each source are independent draws from a stationary non-Gaussian distribution $P(s_j(t))$ [39, 40, 37, 38]. As seen from Figure 4, the sparsity property of time-domain audio signals leads to a sparse distribution, that is a distribution with a "central peak" and "heavy tails" when compared to a Gaussian. An example of such distribution is the generalized exponential distribution

$$P(s_j(t)) \propto \exp(-\beta|s_j(t)|^R) \tag{8}$$

with $0 < R < 2$ and $\beta > 0$. Alternative source models exploiting nongaussianity, nonstationarity and nonwhiteness together or separately have also been proposed [19]. Most ICA algorithms achieve a very good separation performance on determined time-invariant instantaneous mixtures, particularly when the sources are very sparse [37].

Several extensions of ICA have been designed to separate determined time-invariant convolutive mixtures using time-invariant demixing filters[2]. The optimization of time-domain filter coefficients is theoretically feasible [41], but computationally expensive and potentially not robust because of local extrema [42]. Therefore the separation problem is more often addressed in the frequency domain using a STFT representation [43, 15]. Assuming that the STFT coefficients of different sources are independent, a complex demixing matrix $[W_{ji}(f)]_{1\leq i\leq I,1\leq j\leq I}$ can be estimated for each frequency bin f separately using any complex-valued instantaneous ICA algorithm. Source models based on nongaussianity, nonstationarity or nonwhiteness remain valid when transposed to STFT coefficients. For instance, the distribution of source STFT coefficients $P(S_j(n,f))$ in each frequency bin f is generally sparser than the that of the corresponding time-domain signal [36], as shown in Figure 4. Over-determined mixtures can be separated similarly by applying a subspace method in each frequency bin to reduce the number of channels prior to the use of ICA [42].

This frequency-domain separation strategy results in a *permutation problem*, since the estimated source spatial images are arbitrarily ordered in each frequency bin. More information is needed to find the correct permutations so that the source order becomes the same in all frequency bins. One possible approach is to assume that the magnitude

[2] ICA algorithms for instantaneous and convolutive mixtures, including the popular JADE and FastICA algorithms for instantaneous mixtures, are available on the ICA Central web site at `http://www.tsi.enst.fr/icacentral/algos.html`

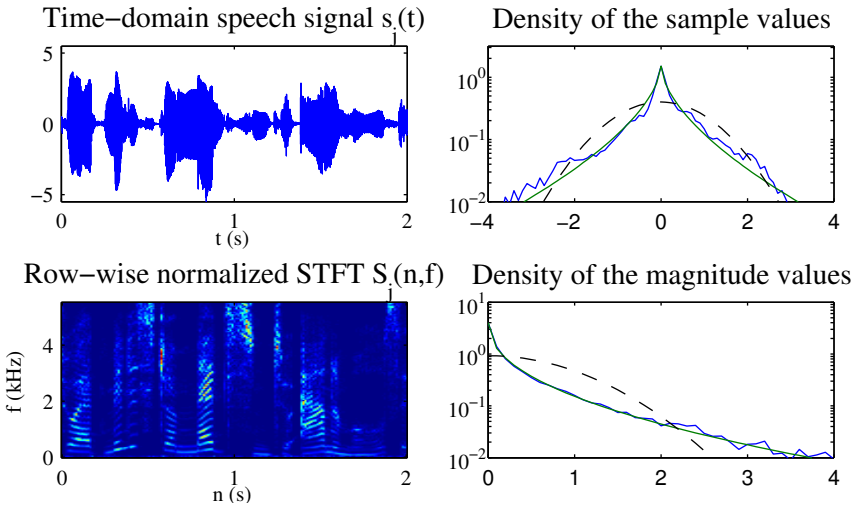

Figure 4: Distributions of a time-domain speech signal and of its magnitude STFT coefficients in each frequency bin (blue curves) compared with generalized exponential distributions of respective parameters $R = 0.60$ and $R = 0.46$ (green curves) and Gaussian distributions of the same variance (dashed curves).

coefficients of each source are correlated between frequency bins and cluster the estimates in order to maximize this correlation [15]. Another approach is to sort the sources according to the estimated delays of arrival between pairs of microphones [44]. These approaches can be formalized in a probabilistic framework by employing coupled source distributions across frequency bins [36] or introducing a spatial penalty term [42]. The algorithm in [45] achieves the best results by combining both approaches, with a reported average SIR of 16 dB on determined mixtures of two static speech sources with 300 ms reverberation time [45]. However performance decreases on time-varying mixtures due to the use of time-invariant filtering [22].

Robustness to source movements can be increased by estimating time-varying demixing matrices using online ICA algorithms [46] or batch ICA algorithms on blocks of time frames [35]. The adaptation speed of these algorithms is limited by the necessity to collect sufficient samples to obtain accurate statistics for the optimized criteria. Adaptation times of several seconds are typically reported for large source movements. Interference rejection can also be improved by post-processing the estimated source spatial images by binary time-frequency masking, which results in a special case of multichannel time-varying filtering and provides a SNR improvement of 1 to 6 dB for two-source mixtures with 300 ms reverberation time [33, 47].

3.2 Directional masking

In the case of under-determined mixtures, the independence assumption used by ICA becomes insufficient to estimate the source STFT coefficients in each frequency bin. More information is needed and sparsity is often exploited to this end. In the limit where the sources are very sparse in the time-frequency domain, one can assume that one source is *dominant* in each time-frequency bin [29, 27]. It is then possible to separate the sources by channel-wise binary masking with the mask being determined from the spatial location information contained in the STFT coefficients of the mixture channels. For stereo mixtures, this information is given by the inter-channel intensity difference (IID)

$$\mathrm{IID}(n,f) = 20\log_{10}\left|\frac{X_2(n,f)}{X_1(n,f)}\right| \tag{9}$$

and the inter-channel phase difference (IPD)

$$\mathrm{IPD}(n,f) = \angle\left(\frac{X_2(n,f)}{X_1(n,f)}\right) \mod 2\pi \tag{10}$$

where $\angle(.)$ denotes the phase of a complex number in $(-\pi, \pi]$. With instantaneous or anechoic mixtures, $\mathrm{IID}(n,f)$ is close to the relative mixing gain $20\log_{10}(|a_{2j}/a_{1j}|)$ for the dominant source j in this time-frequency bin. With

anechoic mixtures, $\mathrm{IPD}(n,f)/2\pi f$ is close modulo $1/f$ to the relative mixing delay $\tau_{2j}-\tau_{1j}$ for the dominant source j. A *phase ambiguity problem* may appear in the upper frequency range: above the frequency $f_{\max}=1/2\tau_{\max}$, where $\tau_{\max}$ is the maximum absolute relative delay, a given value of $\mathrm{IPD}(n,f)$ leads to several possible values of $\tau_{2j}-\tau_{1j}$.

The popular degenerate unmixing estimation technique (DUET) [24] is designed for anechoic mixtures where the mixing filters involve positive gains and fractional delays between -1 and +1 sample, so that the phase ambiguity problem does not arise. The two-dimensional histogram of $(\mathrm{IID}(n,f),\mathrm{IPD}(n,f)/2\pi f)$ is computed and its peaks, which correspond to the relative mixing gains and delays of the sources, are located by a clustering technique. The time-frequency bins within each cluster are then converted into a binary mask $M_j(n,f)$ used to separate the corresponding source spatial image[3]. This algorithm is interpreted in terms of probabilistic source models in [27] by assuming that the source STFT coefficients are drawn from independent uniform distributions under the constraint that exactly one source is active in each time-frequency bin and that the modeling error is Gaussian. A similar algorithm can be derived for instantaneous mixtures using the one-dimensional histogram of $\mathrm{IID}(n,f)$ [48]. An example of application with an instantaneous music mixture is depicted in Figure 5. DUET achieves an average SIR of 16 dB, 12 dB or 6 dB on stereo anechoic speech mixtures with two, three or six sources respectively, but often produces musical noise due to binary masking [27]. Its performance decreases on live recordings, where reverberation or source movements increase the number of directions associated with each source and smear the histogram of IID and IPD, resulting in additional musical noise or inaccurate source direction estimates [27]. For instance, it provides a SIR of 5 dB only for a three-source mixture with 500 ms reverberation time [27].

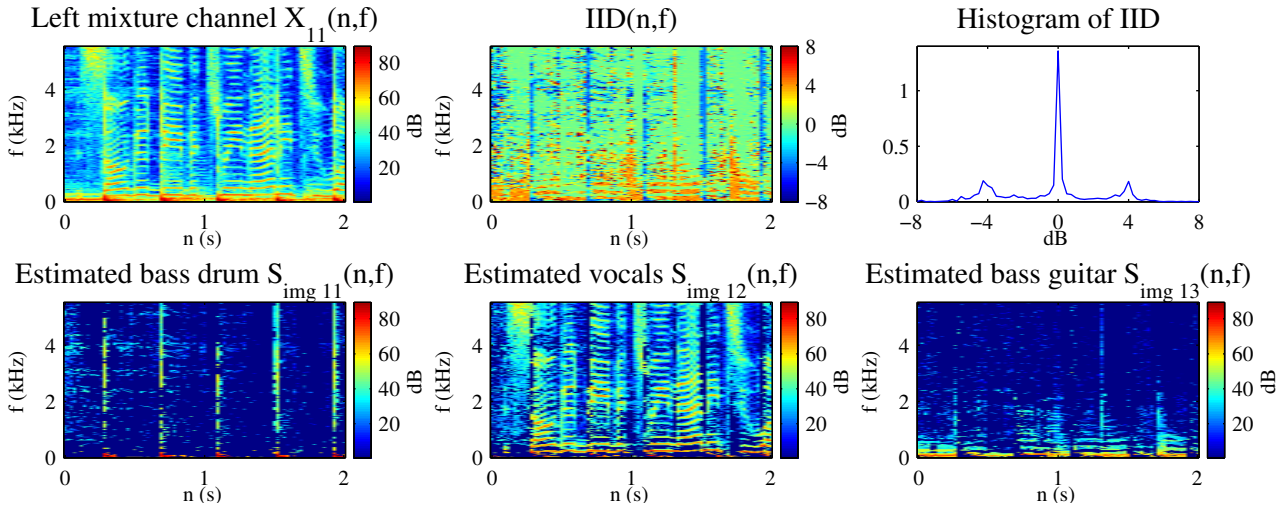

Figure 5: Separation of a three-source instantaneous stereo music mixture via directional binary masking based on IID. The relative mixing gains correspond to IID values of -4, 0 and 4 dB respectively.

The restriction of mixing delays between -1 and +1 sample in DUET, which is rarely satisfied in practice, can be relaxed in the particular case of binaural mixtures, where IID and IPD are dependent on each other and on the sound direction of arrival. In [49], the observed direction is estimated in each time-frequency bin based on the IPD value, where IID is exploited in the upper frequency range to solve the phase ambiguity problem. Then the histogram of directions is computed and the sources are separated by binary masking as above. In [26], an alternative algorithm based on learning the average mapping from IID and IPD to source direction is devised and an average SNR of 12 dB or 4 dB is reported for anechoic binaural mixtures with two or three sources respectively. These binaural separation algorithms suffer the same limitations as DUET regarding reverberation or source movements.

The robustness of the source direction estimates to reverberation can be improved by discarding the time-frequency bins where reverberation dominates when computing the histogram [48]. These bins may be selected based on the observed inter-channel coherence, which is the absolute value of the correlation between $X_1(n,f)$ and $X_2(n,f)$ on a block of time frames along individual frequency bins [29, 48]. This quantity is equal to one when a single point

[3]The clustering technique is not specified in [24]. An approach based on manual detection of the histogram peaks can be easily implemented using the STFT and inverse STFT routines available at `http://www.ee.columbia.edu/~dpwe/resources/matlab/pvoc/`

source is present and it is usually smaller when sounds from different directions overlap. The attenuation of these bins by the masking process can also increase the perceived separation quality [29]. Source movements may be handled by tracking framewise source positions estimates using probabilistic trajectory models [50]. Again, the adaptation speed is limited by the necessity to rely on time continuity priors to reject erroneous local estimates.

The tradeoff between the amount of musical noise and interference may be adjusted by non-binary masking [29]. Nevertheless quality remains insufficient when several sources are active in most time-frequency bins. With time-invariant instantaneous mixtures, a possible approach is to model the source STFT coefficients as independent draws from a sparse distribution $P(S_j(n,f))$ and to estimate them in each frequency bin f by maximizing their probability given the mixture coefficients and a mixing matrix $[A_{ij}]_{1\leq i\leq I, 1\leq j\leq J}$ [34]. This approach is extended to convolutive mixtures in [51, 52] using frequency binwise mixing matrices $[A_{ij}(f)]_{1\leq i\leq I, 1\leq j\leq J}$. These matrices may be estimated by clustering [52] or optimized to maximize the probability of the source coefficients [51]. In certain conditions, this approach is equivalent to multichannel time-varying filtering with I active sources per time-frequency bin where the non-zero rows of the demixing matrix $[W_{ji}(n,f)]_{1\leq i\leq I, 1\leq j\leq J}$ are equal to the pseudo-inverse of the corresponding columns of the mixing matrix [34]. This approach has lead to a SIR of 10 dB and a SNR of 7 dB on a three-source stereo speech mixture with 130 ms reverberation time [52]. However performance appears sensitive to the accuracy of the mixing model (4) and the estimated mixing matrix [52, 28]. A potentially more robust approach is to perform time-varying filtering under the constraint that at most $N < I$ sources are active in each time-frequency bin, where the active sources are estimated assuming a Gaussian modeling error [31, 32]. Note that both approaches also allow the extension of DUET to mixtures with more than two channels.

4 Single-channel separation based on spectro-temporal cues

As an alternative to multichannel source separation methods based on spatial cues, other authors have studied the separation of single-channel mixtures where these cues are absent. In this context, the assumptions of independence and time-frequency sparsity become insufficient and more advanced source models relying on spectro-temporal cues are needed [16]. Existing models typically represent the magnitude of the source STFTs, since phase varies from recording to recording depending on the mixing filters. They provide a generalization of the stationary short-term spectrum models employed for denoising [16] to the case of nonstationary sources. The spectro-temporal cues considered may include discrete structures such as notes and phonemes, along with fundamental frequency, spectral envelope and temporal continuity characteristics. This section reviews three popular models: factorial hidden Markov models, spectral decomposition and monaural computational auditory scene analysis. For simplicity, we drop the channel subscript i in the following and denote by $X(n,f)$ and $[S_{\mathrm{img}\,j}(n,f)]_{1\leq j\leq J}$ the respective STFTs of the single-channel mixture signal and the source spatial images satisfying $X(n,f) = \sum_{j=1}^{J} S_{\mathrm{img}\,j}(n,f)$.

4.1 Factorial hidden Markov models

The simplest source model used for single-channel source separation is perhaps the hidden Markov model (HMM). Assuming that the short-term log-power spectrum of each source spatial image is modelled by an HMM with Gaussian observation distributions with shared diagonal covariance, this gives [16]

$$P(\log|S_{\mathrm{img}\,j}(n,f)|^2) = \mathcal{N}(\log|S_{\mathrm{img}\,j}(n,f)|^2; \Psi_{h_j(n)}(f), \sigma_j^2(f)) \tag{11}$$

where $h_j(n)$ is a hidden state belonging to a finite set $\mathcal{H}_j$ called the *state space*, $\Psi_{h_j(n)}(f)$ is the mean log-power spectrum associated with this state and $\mathcal{N}(.;\mu,\sigma^2)$ is the univariate Gaussian density with mean μ and variance σ^2. Each hidden state may represent a particular phoneme at a given fundamental frequency or a particular chord. The state sequence $[h_j(n)]_{1\leq n\leq N}$ follows a first order Markov prior defined by the multinomial distributions $P(h_j(1))$ and $P(h_j(n+1) \mid h_j(n))$, which model the average duration of each state and the probability of transition from one state to another. A mixture of several sources may be represented by combining the source HMMs into a single mixture HMM whose hidden states are J-uples of the form $h(n) = (h_1(n), \ldots, h_J(n))$. The mixture state space is then the Cartesian product $\mathcal{H}_1 \times \cdots \times \mathcal{H}_J$ of the source state spaces and its size grows exponentially with the number of sources. When the sources are assumed to be independent, the prior probability of the mixture state sequence

factorizes as the product of the prior probabilities of the source state sequences

$$P([h(n)]_{1\leq n\leq N}) = \prod_{j=1}^{J} P(h_j(1)) \prod_{n=2}^{N} P(h_j(n+1) \mid h_j(n)) \tag{12}$$

and the resulting model is termed a *factorial* HMM. Under this independence assumption, the power spectrum of the mixture is equal on average to the sum of the source power spectra and the distribution $P(\log|X(n,f)|^2 \mid h(n))$ of the short-term log-power spectrum for a given mixture state $h(n)$ can be approximated as a Gaussian whose mean and variance are non-trivial functions of the means and variances of the underlying source observation distributions [53]. When the latter share the same variance $\sigma^2(f)$, a simpler approximation is given by the *log-max* formula [23]

$$P(\log|X(n,f)|^2 \mid h(n)) \approx \mathcal{N}\Big(\log|X(n,f)|^2; \max_{j=1,\ldots,J} \Psi_{h_j(n)}(f), \sigma^2(f)\Big). \tag{13}$$

The use of factorial HMMs for single-channel speech separation follows four steps described in [16, 23]. Firstly, an individual HMM is trained for each source beforehand on solo (single-source) training examples using the expectation-maximization (EM) algorithm. Secondly, the most probable mixture state sequence $[\widehat{h}(n)]_{1\leq n\leq N}$ is inferred using the Viterbi algorithm along with some heuristics to avoid testing improbable sequences. Thirdly, the log-power spectrum of each source spatial image is derived from the source state sequences by $\log|\widehat{S}_{\mathrm{img}\,j}(n,f)|^2 = \Psi_{\widehat{h}_j(n)}(f)$. Finally, the corresponding signals are computed via time-frequency masking either by attributing each time-frequency bin to the loudest source [23] or by deriving non-binary masks from the Wiener filtering formula [16][4]. This algorithm has been claimed to achieve a good performance on a mixture of male and female speech [23], as illustrated in Figure 6. However time-varying permutation errors may appear for mixtures of speakers with similar fundamental frequency range and timbre. Performance can be greatly improved in this situation by introducing constrained *language models* in the source HMMs to account for the long-term structure of speech [53].

For music mixtures, the application of factorial HMMs raises some difficulties that are not encountered with speech [54]. The natural state space underlying music sources is often much larger than for speech sources because of their wider fundamental frequency and intensity range and their ability to produce chords. Consequently, HMMs using a small number of states provide a poor separation performance because they model the sources too coarsely. However HMMs using many states may also provide a poor performance due to *overfitting*: the amount of data available to train the parameters of each state becomes so small that the models do not generalize well to data outside the training set. In [54], a solution to this problem is provided in the case of mixtures of singing voice and accompaniment music by segmenting the mixture manually. The HMM representing accompaniment music is trained on the segments of the mixture containing accompaniment music only, while the HMM representing the singing voice is first trained on a general corpus of singing voice and then adapted to the mixture using the segments containing both singing voice and accompaniment music. This approach results in a SNR of 10 dB, compared with a SNR of 5 dB only using generic voice and accompaniment models [54].

4.2 Spectral decomposition

Another way to model audio sources is to approximate the short-term power spectrum of each source spatial image by a weighted sum of *basis spectra* [55]. For example, the observed spectrum of a musical chord may be decomposed as a weighted sum of note spectra, which may be in turn decomposed as weighted sums of basis spectra with different spectral envelopes, but with the same fundamental frequency if periodic. Assuming that the power spectrum of the single-channel mixture signal is equal on average to the sum of the source power spectra and that the modeling error is Gaussian with diagonal nonstationary covariance, this gives

$$P(|X(n,f)|^2) = \mathcal{N}\left(|X(n,f)|^2; \sum_{j=1}^{J}\sum_{k=1}^{K_j} e_{jk}(n)\Phi_{jk}(f), \sigma^2(n,f)\right) \tag{14}$$

[4]These steps can be implemented using the HMM toolbox at `http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html` along with the above mentioned STFT and inverse STFT routines.

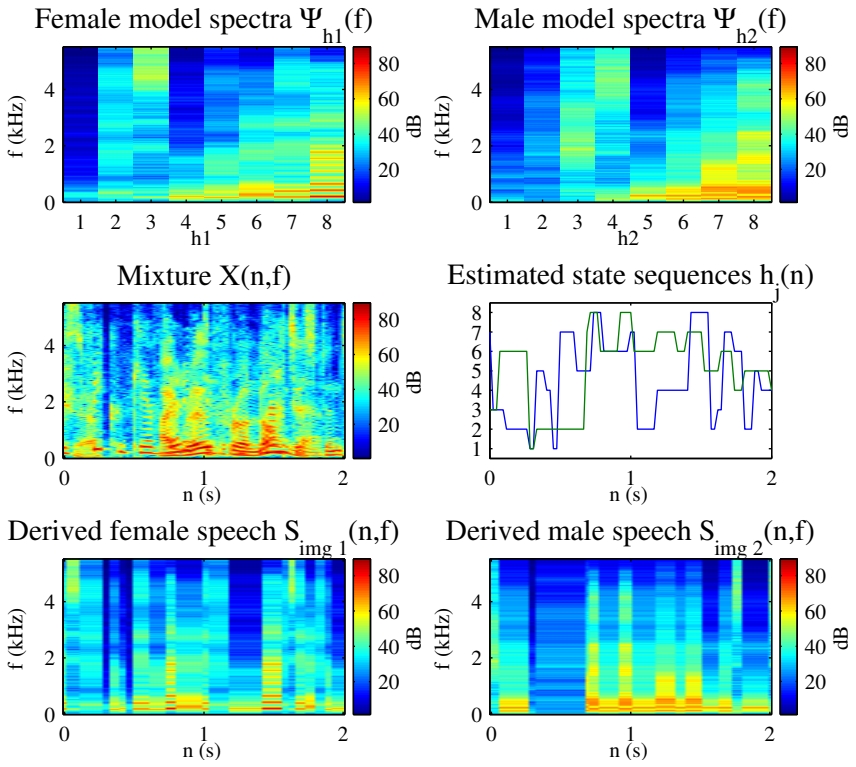

Figure 6: Separation of a two-source single-channel speech mixture using a factorial HMM with 8 states per source. The estimated state sequences for the female and the male source are plotted in blue and green respectively. The derived source magnitude STFTs are obtained directly from the state spectra and the estimated state sequences prior to the application of time-frequency masking. The STFTs of the resulting source signals and the solo training data for each source are not shown.

where $[\Phi_{jk}(f)]_{1\leq k\leq K_j}$ and $[e_{jk}(n)]_{1\leq k\leq K_j}$ are respectively the basis spectra and the time-varying weights for source j. Generally, weight sequences from the same source show some statistical dependencies: for example, notes composing a chord are activated at the same time. However weights from different sources are assumed to be independent. This spectral decomposition model is particularly efficient for music, since it allows a huge reduction of the model size: only a few basis spectra per note are needed to represent accurately a given source instead of several states per chord in the case of an HMM. Further size reduction may be achieved by assuming that all the periodic basis spectra of each source are obtained by translation of a single mother spectrum on the log-frequency axis [56] or by stacking successive time frames into short spectrograms and representing each spectrogram as a weighted sum of basis spectrograms capturing short-term nonstationarity [57].

Spectral decomposition has mostly been applied to the separation of music mixtures via a three-step data-driven approach. The most probable basis spectra are first estimated from the mixture signal using one of a range of probabilistic models, under the constraint that both the basis spectra and the time-varying weights are nonnegative. In certain conditions, this constraint suffices to obtain relevant basis spectra with a fixed variance σ^2 using a nonnegative matrix factorization (NMF) algorithm [58]. Alternatively, similar algorithms may be employed with more complex models involving sparsity or temporal continuity priors on the weight sequences [59, 56]. The modeling accuracy may be increased in low-power time-frequency regions by adjusting the variance $\sigma^2(n,f)$ in each time-frequency bin according to human audition [57]. Once the basis spectra have been estimated, groups of spectra corresponding to different sources are built by clustering together the weight sequences with the most similar distributions [55] or the basis spectra associated with the same mother spectrum [56]. Finally, the power spectrum of each source spatial image is derived by $|\widehat{S}_{\text{img}\,j}(n,f)|^2 = \sum_{k=1}^{K_j} \widehat{e}_{jk}(n)\widehat{\Phi}_{jk}(f)$ and the corresponding time-domain signals are extracted by time-frequency masking[5]. This approach has shown to provide good separation results for synthetic music excerpts

[5]Simpler algorithms based on manual clustering of the basis spectra can be implemented using the STFT and inverse STFT routines

[57]. The separation of a drum loop is illustrated in Figure 7.

Data-driven clustering appears more difficult to apply to real-world music signals, where instruments often play synchronously and note spectra vary in a way that cannot be represented by the translation of a single mother spectrum. An alternative approach is to classify the estimated basis spectra into fixed source classes using supervised classifiers. This has been used for the separation of drums from non-percussive instruments [59] and vocals from accompaniment [60]. However some interference generally remains, even when the clusters are chosen manually, because some of the basis spectra represent notes from different sources [57]. This issue can sometimes be addressed by prior learning of the basis spectra on solo excerpts of each of the sources. In [61], a SIR of 15 dB is reported for the separation of a synthetic mixture of cello and drum sources using basis spectra trained on a different part of the same source signals. This suggests that spectral decomposition can provide a better separation performance than factorial HMMs for music mixtures, when the model parameters are trained in a supervised fashion for both methods.

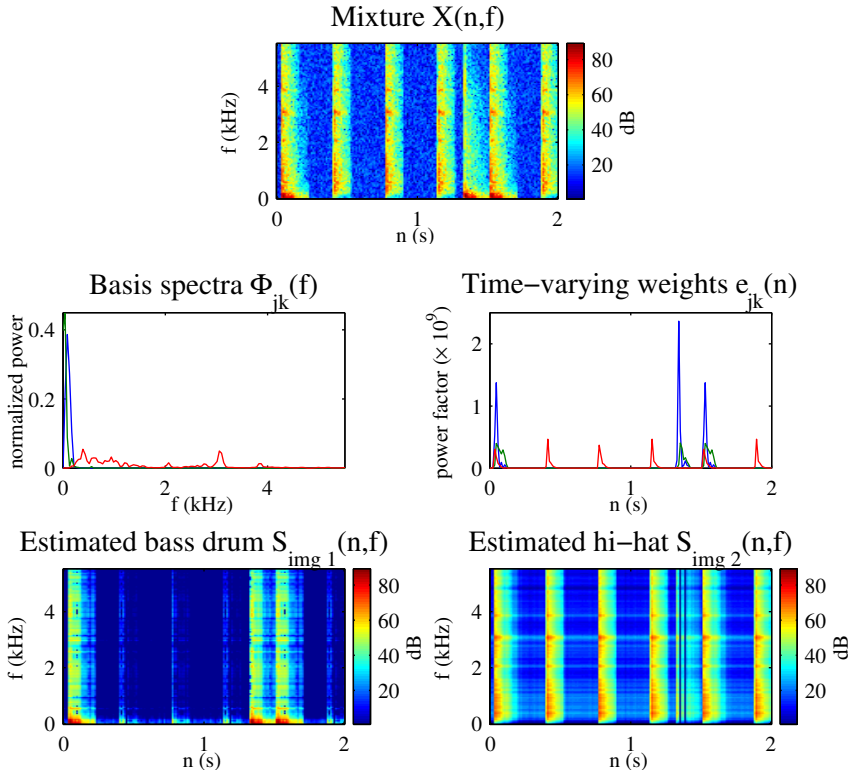

Figure 7: Separation of a two-source synthetic drum loop using spectral decomposition via NMF. Bass drum is modeled by the blue and green components and hi-hat by the red component. The derived source magnitude STFTs are obtained directly from the basis spectra and the time-varying weights prior to the application of time-frequency masking. The STFTs of the resulting source signals are not shown.

4.3 Monaural computational auditory scene analysis

Historically, the motivation behind spectral decomposition was not to improve the modeling of music sources but to provide a statistical model of auditory organization, where the observed short-time spectrum is explained by a small number of objects at each instant, each modeled by a stationary spectrum [55]. Audition is known to segregate sources within a complex scene based on a similar redundancy reduction process [62]. However, auditory objects generally have nonstationary spectra, and other cues are taken into account to group time-frequency regions into objects and stream these objects into sources. Listening experiments have led to five *grouping* and *streaming* rules called proximity, similarity, continuity, closure and common fate [62], named from *Gestalt* psychology theory. These rules state for example that a set of sinusoidal partials constituting a periodic object tend to have harmonic frequencies, a smooth

mentioned above and the NMF update rules provided in [58].

spectral envelope, similar onset and offset times and correlated amplitude and frequency variations. Computational auditory scene analysis (CASA) aims to analyze speech and music signals by implementing several of these rules. Some principles of auditory organization and the derived CASA algorithms are reviewed in *e.g.* [63, 64].

Early CASA algorithms focus on the extraction of objects from *monaural* (single-channel) signals and contain four successive processing stages [65, 25]. Firstly, the mixture signal is transformed into a front-end representation which is easier to process. Most often, this is done by splitting the signal into several subbands using a perceptually motivated filterbank and computing the autocorrelation function of the absolute value of each subband signal on short time frames, leading to a three-dimensional representation known as *correlogram*. For simplicity, this representation is sometimes replaced by the STFT magnitude [66]. Secondly, a collection of sinusoidal partials is extracted from the front-end representation, for example by locating and tracking over time the peaks of the autocorrelation function or the magnitude spectrum. Thirdly, these partials are organized iteratively into objects by applying the grouping rules in a fixed order to the longest remaining partial. Finally, the objects are extracted by binary masking.

This data-driven approach is fast but lacks some robustness [66, 67]. Indeed, when a given subband contains sinusoidal partials from different sources, the partials corresponding to low power sources may be either not detected at all or transcribed with erroneous onset times that induce grouping errors due to the rigid evidence integration process. The algorithm in [67] solves this problem by exploiting advanced object models to correct the obscured parameters in a prediction-driven fashion. More precisely, three types of objects called "wefts", "transients" and "noise clouds" are defined. For example, wefts are periodic objects made of harmonic sinusoidal partials with equal onset and offset times and constant spectral envelope. Inference is carried out by scanning the time frames in ascending order and testing several competing hypotheses (*i.e.* sets of objects) within a blackboard architecture, as shown in Figure 8. The set of hypotheses grows iteratively by prolongating, resuming or creating objects based on harmonicity and onset cues. Each hypothesis is associated with a score measuring the modeling error between the underlying object models and the front-end representation, and the best scored hypothesis is selected in the end. This approach may be implemented similarly in a probabilistic framework, where the objects and the modeling error are modelled by prior distributions and the hypothesis with the best posterior probability is selected [66].

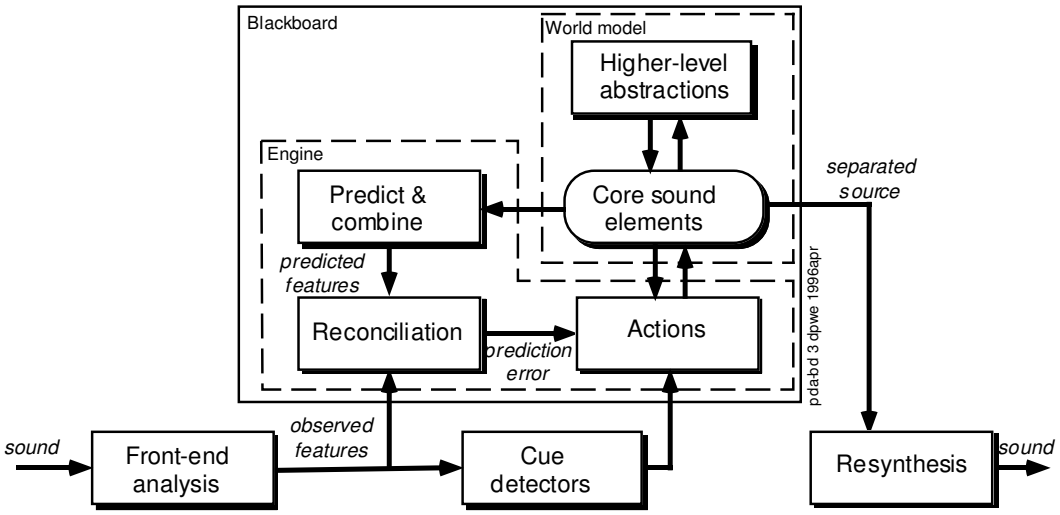

Figure 8: Schema of a prediction-driven CASA system (from [67], with permission).

While early CASA algorithms did not consider the streaming of objects into sources, more recent algorithms have addressed the source separation problem by scoring competing streaming hypotheses according to additional language or musicological rules. One algorithm suited to mixtures containing one speech source and other non-speech sources searches for the sequence of objects that results in the most probable word sequence given an HMM speech model trained on clean speech data [68]. This yields a good separation of a speech source mixed with industrial noise. A similar approach can be devised for mixtures containing several concurrent speech sources by replacing the single-source HMM with a multi-source factorial HMM. However the problem becomes more difficult in this case and speaker-specific models may be needed. Other CASA algorithms proposed for music mixtures exploit musical knowledge and timbre features learnt on solo excerpts to cluster periodic objects into instruments [66, 69]. These algorithms

potentially improve the instrument segregation performance compared to HMM and spectral decomposition methods since they use advanced timbre features such as onset duration and frequency modulation in addition to the spectral envelope feature. Moreover, they provide a more natural model for notes with nonstationary fundamental frequency.

5 Multichannel separation based on hybrid models

The main limitation of the single-channel source separation methods described above is that the sources must have noticeably different fundamental frequency ranges or timbres. When dealing with a mixture of male speakers or a violin duo for instance, the sources can be well separated on short time frames but time-varying permutation errors are likely to occur. Streaming can be improved using constrained language or musicological models, at the expense of an increased computational cost. Nevertheless single-channel mixtures where the sources exhibit synchronous temporal variations remain very difficult to separate, even by the human auditory system [62]. Since spatial properties are believed to play an important role for long-term auditory streaming [62], some authors have proposed to adapt single-channel source models to multichannel signals by jointly modeling spatial and spectro-temporal cues. These hybrid models also potentially address some limitations of conventional multichannel source separation methods on reverberant or under-determined mixtures.

Factorial HMMs are extended to binaural mixtures in [70] by exploiting the IID cues (9) introduced previously for directional masking. Given the underlying mixture state, the short-term log-power spectra of the two mixture channels are modelled by independent Gaussian distributions taking into account the expected value of IID for the hypothesized source directions. The source power spectra are estimated by maximizing the joint probability of the mixture state sequence and the source directions, and the corresponding time-domain signals are derived by time-frequency masking. Promising results are reported for anechoic speech mixtures assuming perfect knowledge of the mixing filters as a function of the source directions. A different extension of factorial HMMs is proposed in [71] for the extraction of a target source from a multichannel convolutive mixture by time-invariant filtering. This algorithm outperforms ICA on over-determined mixtures of two speakers with 200 ms reverberation time when the sentence pronounced by each speaker is known.

A multichannel spectral decomposition algorithm for instantaneous mixtures is proposed in [72]. The short-term power spectra of the mixture channels are modelled by summing single-channel basis spectra with time-varying weights and channel-specific mixing gains. The most probable basis spectra and mixing gains are estimated via a modified NMF algorithm, assuming independent modeling errors on different channels. The basis spectra are then grouped into sources according to their mixing gains and time-frequency masking is performed. Good separation results are claimed for a mixture of piano and drums.

CASA algorithms can easily incorporate spatial cues by adding a spatial proximity rule in the hypothesis scoring process. The binaural CASA algorithm proposed in [73] selects the harmonic partials forming a periodic object and the objects forming a stream based in particular on the distance between their directions of arrival computed from IID and IPD cues (9)-(10). When applied to time-invariant anechoic speech mixtures, this algorithm achieves a better separation performance than its single-channel counterpart, which exploits continuity and frequency proximity only. This approach has been extended to mixtures with more than two channels in [74]. A similar algorithm has been proved to increase the separation performance for anechoic music mixtures by using spatial proximity and spectral envelope proximity rules jointly instead of using one of these rules only [75].

The algorithm introduced in [76] combines spectral decomposition, factorial HMMs and IPD cues into a single probabilistic model for music mixtures. The short-term power spectrum of each source is represented by a weighted sum of instrument-specific note spectra learnt beforehand for different fundamental frequencies and each frequency is associated with a binary activity state modeled by a Markov prior. The power spectrum of the mixture is assumed to be equal to the sum of the source power spectra multiplied by the magnitude response of the mixing filters, and the observed IPD is expressed as a function of the source power spectra and the source directions. The most probable activity states and time-varying weights are estimated by beam search and the source spatial images are computed by channel-wise time-frequency masking. This algorithm provides a SNR of 14 dB on stereo two-source music mixtures with 800 ms reverberation time relying on prior knowledge of the instrument classes and the source directions, outperforming ICA and directional masking by more than 10 dB [76]. The SNR decreases by 1 dB when the source directions are misestimated by 5 degrees.

Simpler hybrid models that do not attempt to model the full spectro-temporal structure of the sources have also been proposed. For instance, the separation of under-determined instantaneous mixtures via time-varying filtering is addressed by maximizing the correlation between the magnitude of each source in harmonically related frequency bins in [77] and by using a temporal continuity model in each frequency bin in [78]. These models improve the separation performance over directional masking in certain conditions at a small computational cost. Overall, hybrid models remain a fairly recent approach to audio source separation and existing hybrid algorithms seem to have been designed as proofs of concept rather than optimized for speed or performance.

6 Discussion

Despite their different historical backgrounds, the model-based source separation methods reviewed in this article share many common features. The summarizing of these features leads to a better understanding of their behavior and their limitations for real-world audio signals.

6.1 Summary classification

As seen throughout the article, model-based audio source separation algorithms are usually based on a parametric approach, where separation is conducted by estimating the parameters of a given separating function using a certain model. This allows the two-way classification of source separation algorithms according to the chosen family of separating functions on the one side and the exploited signal model on the other side.

Three families of separating functions may be distinguished: multichannel time-invariant filtering relies on spatial diversity, channel-wise time-frequency masking on time-frequency diversity, and multichannel time-varying filtering on either of these assumptions. While a majority of convolutive ICA algorithms separate the source spatial images by multichannel time-invariant filtering, most other algorithms do so by channel-wise time-frequency masking. Multichannel time-varying filtering is employed by a few algorithms.

The underlying signal models appear more different for historical reasons. For instance, ICA and CASA were initially introduced in the fields of neural networks [6] and computational perception [7] respectively, while DUET was designed as a practical source separation procedure [24]. These methods have since been reformulated in terms of probabilistic signal models [37, 66, 27], allowing their comparison with factorial HMMs and spectral decomposition. Many recent algorithms adopt a probabilistic framework, since it provides both a natural way of designing hybrid signal models and also a unique separation objective, specifically the maximization of the posterior probability of the sources. The assumptions made by different models are summarized in Table 1. It can be seen that most models rely on the independence of the sources and share some additional assumptions. Some directional masking models assume a fixed number of inactive sources in each time-frequency bin, which introduces a dependence between the sources, however the active sources are still assumed to be independent.

Interestingly, the choice of the family of separating functions is not always dictated by the signal model and may rely on different assumptions about the sources. For example, any model of the source magnitude STFTs may be used for separation either by channel-wise time-frequency masking or by multichannel time-invariant filtering, as shown with factorial HMMs [16, 71].

6.2 Performance comparison

The above two-way classification is also relevant to study the performance of source separation algorithms. Indeed the performance of a given algorithm may be limited due either to the constraints inherent to the family of separating functions or to the lack of information in the signal model. Additional limitations due to the use of a non-perfect optimization algorithm may also arise.

The performance upper bounds of multichannel time-invariant filtering and channel-wise time-frequency masking have been studied separately in [22, 27] and compared more recently in [28]. Multichannel time-invariant filtering appears to perform better than channel-wise time-frequency masking for determined or over-determined mixtures, but similarly or worse for under-determined mixtures, particularly with non-binary masks. Moreover its performance decreases more quickly in reverberant conditions. This validates the choice of time-frequency masking as opposed to

Table 1: Summary of the source properties (rows) exploited by different categories of models (columns). Particular models may rely only on some of the properties listed for the corresponding category.

	Convolutive ICA	Directional masking	Factorial HMM	Spectral decomposition	Monaural CASA	Hybrid models
Independence	✓	✓	✓	✓	✓	✓
Number of inactive sources		✓				✓
Spatial position	✓	✓				✓
Sparsity or nongaussianity	✓	✓		✓		✓
Correlation across frequency	✓			✓	✓	✓
Spectral envelope and timbre			✓	✓	✓	✓
Harmonicity			✓		✓	✓
Language or music rules			✓		✓	✓

time-invariant filtering by almost all algorithms aiming to separate under-determined mixtures. Time-varying filtering with a generalized disjoint orthogonality constraint outperforms both time-invariant filtering and time-frequency masking for determined and under-determined mixtures, but is more sensitive to the accurate estimation of the mixing parameters. See [28] for exact SNR values.

The performance of the various signal models is more difficult to assess, since no wide-scale evaluation of source separation algorithms has yet been conducted. The SNR and SIR values reported in the previous sections have been computed on different datasets for a few algorithms only and cannot be directly compared. From a qualitative point of view, models based on spatial cues cannot discriminate sources from the same direction, while models based on spectro-temporal cues cannot easily separate sources with similar pitch range and timbre. Therefore solving the source separation problem in a general context implies the joint exploitation of both types of cues. This claim is supported by the numerical performance of some hybrid models [73, 74, 76] and experimental evidence regarding the human auditory system [5, 74].

6.3 Relation to beamforming

As mentioned in the introduction, model-based source separation has been developed as an alternative to beamforming, which extracts a target signal based on its direction of arrival and on the relative microphone positions without any prior knowledge about the signals themselves [1, 2]. The target direction can be either provided manually or estimated using a cross-correlation-based algorithm. Beamforming algorithms are classified into two categories: fixed beamformers enhance sounds from the target direction regardless of the spatial distribution of interference and adaptive beamformers attempt to minimize the energy of interference without distorting sounds from the target direction. These algorithms are most often implemented via multichannel time-invariant or time-varying filtering. In practice, adaptive beamformers are optimized on the time-frequency bins where the target is inactive, which can be estimated by comparing the input and output energy of a fixed beamformer pointing towards the target direction [79].

The performance achievable by beamforming is very good when a large microphone array is available [2]. However it decreases when the number of microphones is small [3, 4], partly because the upper performance bound of multichannel filtering decreases [21] but also because the target direction and the time-frequency bins where the target is inactive are estimated less accurately. By contrast, convolutive ICA can adapt the demixing filters even when all the sources are active simultaneously and implicitly estimates the source directions maximizing the independence of the sources. Also it does not require prior knowledge of the microphone positions, which makes it robust to array deformations. Adaptive beamforming and convolutive ICA with nonstationary Gaussian source models are compared in more details theoretically in [80] and experimentally in [42], where convolutive ICA is shown to result in a superior performance with three sources and eight microphones.

Beamforming algorithms are also sometimes implemented via time-frequency masking and then called nonlinear beamformers. Simple directional masking algorithms such as DUET can be seen as nonlinear beamformers, with the difference that prior knowledge of the microphone positions is not necessary. The performance of these algorithms decreases when the estimated source directions are not accurate. By contrast, some advanced directional masking al-

gorithms implicitly estimate the source directions maximizing the probability of the sources under a sparse distribution model [51].

The knowledge of the relative microphone positions can also be exploited within model-based algorithms to segregate the sources based on explicit directions of arrival instead of relative delays and intensities between pairs of microphones. A hybrid model using this information is proposed in [74].

6.4 Concluding remarks and future challenges

Overall, it is not possible to recommend one model-based algorithm as the best for all situations. Directional masking allows real-time low-delay separation of multichannel mixtures, but often produces musical noise [24]. Convolutive ICA, possibly followed by time-frequency masking, generally performs better on determined and over-determined mixtures, at the expense of a longer processing delay [47]. Hybrid methods exploiting spatial and spectro-temporal cues jointly may reduce musical noise on under-determined or reverberant mixtures, but must be applied offline [76]. Single-channel methods seem to provide a lower performance and are rarely applied to multichannel mixtures, except when the mixing filters may have different properties than natural room impulse responses, such as in recent music CDs [54]. Some single-channel and hybrid methods necessitate prior information, such as the categorization of the sources as speech or music or solo training data for each speaker or instrument. No method so far provides a sufficient quality on real-world signals for high-fidelity applications involving the creation of a music piece or a movie soundtrack from the separated sources. This is particularly evident for mixtures involving a large number of sources, reverberation or source movements. Some recent methods attempt to tackle these difficult conditions.

Under-determined mixtures are generally separated in the STFT domain via channel-wise time-frequency masking or multichannel time-varying filtering under a generalized disjoint orthogonality constraint. The resulting performance upper bound decreases when many sources are present, since the number of active sources in each time-frequency bin becomes larger [27]. This issue could be overcome by applying the separating functions in a different domain where the overlap between sources is smaller or by building source waveform models achieving a higher spectro-temporal resolution. Promising results have already been obtained using warped-frequency filterbanks [81] and harmonic-plus-noise waveform models [82]. Adaptive time-frequency representations could also be investigated.

Separation methods based on spatial cues often perform poorly in reverberant environments, since the performance upper bound of multichannel time-invariant filtering is reduced and the association of each direction of arrival with a given source becomes ambiguous. Hybrid methods exploiting spatial and spectro-temporal cues jointly can improve performance in this context, as shown in [76]. Performance remains limited nevertheless since reverberation results in a temporal smearing of fundamental frequency and a modification of the spectral envelope. These spectro-temporal cues could be estimated more accurately by applying dereverberation filters on selected parts of the mixture signal. Good preliminary results have been demonstrated using knowledge of the optimal filters in [83].

Few model-based methods have been proposed so far for the separation of multichannel time-varying mixtures. Most of them rely on the tracking of demixing matrices [46, 35] or source directions [50] estimated from spatial cues on each time frame. Long adaptation times are necessary to smooth out erroneous estimates. Moreover permutations of the separated sources may occur when the source trajectories cross each other or when sources move during silence. Spectro-temporal cues could help improve performance by providing more precise source direction estimates leading to shorter adaptation times and allowing tracking based on timbre cues. Promising results have been reported on anechoic mixtures using a hybrid model of source trajectories and short-term spectra based on IID cues [70].

In conclusion, model-based methods have already achieved very promising results for the separation of mixture signals with few channels. However some types of signals remain difficult to separate due to one or more issues including a large number of sources, reverberation or source movements. The design of new methods addressing these issues is perhaps the greatest challenge for future research.

Acknowledgments

Emmanuel Vincent, Maria G. Jafari and Samer A. Abdallah are funded by EPSRC grants GR/S75802/01, GR/S85900/01 and GR/S82213/01 respectively. Mike E. Davies acknowledges support for his position from the Scottish Funding Council and for their support of the Joint Research Institute with the Heriot-Watt University as part of the Edinburgh

Research Partnership. The authors wish to thank the anonymous reviewers for very useful comments on early versions of this article, as well as Daniel Ellis, Katy Noland and Andrew Nesbit.

References

[1] B. D. van Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[2] M. S. Brandstein and D. B. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, 2001.

[3] J. E. Greenberg and P. M. Zurek, "Evaluation of an adaptive beamforming method for hearing aids," *Journal of the Acoustical Society of America*, vol. 91, no. 3, pp. 1662–1676, 1992.

[4] M. E. Lockwood, D. L. Jones, R. C. Bilger, C. R. Lansing, W. D. O'Brien, B. C. Wheeler, and A. S. Feng, "Performance of time- and frequency-domain binaural beamformers based on recorded signals from real rooms," *Journal of the Acoustical Society of America*, vol. 115, no. 1, pp. 379–391, 2004.

[5] J. F. Culling, K. I. Hodder, and C. Y. Toh, "Effects of reverberation on perceptual segregation of competing voices," *Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2871–2876, 2003.

[6] J. Hérault and C. Jutten, "Space or time adaptive signal processing by neural network models," in *Proc. American Institute of Physics (AIP) Conf.*, Snowbird, UT, 1986, pp. 206–211.

[7] M. Weintraub, "A theory and computational model of auditory monaural sound separation," Ph.D. dissertation, Dept. of Eletrical Engineering, Stanford University, 1985.

[8] L. Deng and D. O'Shaughnessy, *Speech Processing: A Dynamic and Optimization-Oriented Approach*. Marcel Dekker, Taylor & Francis, 2003.

[9] D. E. Hall, *Musical Acoustics, 3rd Edition*. Brooks Cole, 2001.

[10] B. Gygi, G. R. Kidd, and C. S. Watson, "Spectral-temporal factors in the identification of environmental sounds," *Journal of the Acoustical Society of America*, vol. 115, no. 3, pp. 1252–1265, 2004.

[11] B. Bartlett and J. Bartlett, *Practical Recording Techniques, 4th Edition : the Step-by-step Approach to Professional Recording*. Focal Press, 2005.

[12] H. Kuttruff, *Room Acoustics, 4th Edition*. Spon Press, 2000.

[13] J.-F. Cardoso, "Multidimensional independent component analysis," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1998, pp. IV–1941–1944.

[14] P. D. O'Grady, B. A. Pearlmutter, and S. T. Rickard, "Survey of sparse and non-sparse methods in source separation," *Int. Journal of Imaging Systems and Technology*, vol. 15, pp. 18–33, 2005.

[15] S. Ikeda and N. Murata, "An approach to blind source separation of speech signals," in *Proc. Int. Conf. on Artificial Neural Networks (ICANN)*, 1998, pp. 761–766.

[16] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proceedings of the IEEE*, vol. 80, no. 10, pp. 1526–1555, 1992.

[17] A. Mansour, M. Kawamoto, and N. Ohnishi, "A survey of the performance indexes of ICA algorithms," in *Proc. IASTED Int. Conf. on Modelling, Identification and Control (MIC)*, 2002, pp. 660–666.

[18] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[19] H. Buchner, R. Aichner, and W. Kellermann, "Blind source separation for convolutive mixtures: a unified treatment," in *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, Y. Huang and J. W. Benesty, Eds. Kluwer, 2004, pp. 255–294.

[20] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 36, no. 2, pp. 145–153, 1988.

[21] R. Mukai, S. Araki, H. Sawada, and S. Makino, "Evaluation of separation and dereverberation performance in frequency domain blind source separation," *Acoustical Science and Technology*, vol. 25, no. 2, pp. 119–126, 2004.

[22] R. V. Balan, J. P. Rosca, and S. T. Rickard, "Robustness of parametric source demixing in echoic environments," in *Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA)*, 2001, pp. 144–148.

[23] S. T. Roweis, "One microphone source separation," in *Advances in Neural Information Processing Systems (NIPS 13)*, 2001, pp. 793–799.

[24] A. N. Jourjine, S. T. Rickard, and Ö. Yılmaz, "Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2000, pp. V–2985–2988.

[25] G. J. Brown and M. P. Cooke, "Computational auditory scene analysis," *Computer Speech and Language*, vol. 8, pp. 297–336, 1994.

[26] N. Roman, D. L. Wang, and G. J. Brown, "Speech segregation based on sound localization," *Journal of the Acoustical Society of America*, vol. 114, no. 4, pp. 2236–2252, 2003.

[27] Ö. Yılmaz and S. T. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[28] E. Vincent, R. Gribonval, and M. D. Plumbley, "Oracle estimators for the benchmarking of source separation algorithms," Queen Mary, University of London, Tech. Rep. C4DM-TR-06-03, july 2006. [Online]. Available: http://www.elec.qmul.ac.uk/people/markp/2006/VincentGribonvalPlumbley06-tr.pdf

[29] B. Kollmeier, J. Peissig, and V. Hohmann, "Real-time multiband dynamic compression and noise reduction for binaural hearing aids," *Journal of Rehabilitation Research and Development*, vol. 30, no. 1, pp. 82–94, 1993.

[30] N. Virag, "Single-channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. on Speech and Audio Processing*, vol. 7, no. 2, pp. 126–134, 1999.

[31] L. Vielva, D. Erdoğmuş, and J. C. Príncipe, "Underdetermined blind source separation using a probabilistic source sparsity model," in *Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA)*, 2001, pp. 675–679.

[32] J. P. Rosca, C. Borss, and R. V. Balan, "Generalized sparse signal mixing model and application to noisy blind source separation," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2004, pp. III–877–880.

[33] D. Kolossa and R. Orglmeister, "Nonlinear postprocessing for blind speech separation," in *Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA)*, 2004, pp. 832–839.

[34] M. Zibulevsky, B. A. Pearlmutter, P. Bofill, and P. Kisilev, "Blind source separation by sparse decomposition in a signal dictionary," in *Independent Component Analysis : Principles and Practice*. Cambridge Press, 2001, pp. 181–208.

[35] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Blind source separation for moving speech signals using blockwise ICA and residual crosstalk subtraction," *IEICE Trans. Fundamentals*, vol. E87-A, no. 8, pp. 1941–1948, 2004.

[36] N. Mitianoudis and M. E. Davies, "Audio source separation of convolutive mixtures," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 5, pp. 489–497, 2003.

[37] J.-F. Cardoso, "Blind source separation: statistical principles," *Proceedings of the IEEE*, vol. 9, no. 10, pp. 2009–2025, 1998.

[38] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. Wiley, 2001.

[39] P. Comon, "Independent component analysis - a new concept?" *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[40] A. J. Bell and T. J. Sejnowski, "An information maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, pp. 1129–1159, 1995.

[41] K. Torkkola, "Blind separation of convolved sources based on information maximization," in *Proc. IEEE Workshop on Neural Networks for Signal Processing (NNSP)*, 1996, pp. 423–432.

[42] L. C. Parra and C. V. Alvino, "Geometric source separation : merging convolutive source separation with geometric beamforming," *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 6, pp. 352–362, 2002.

[43] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, pp. 21–34, 1998.

[44] M. Z. Ikram and D. R. Morgan, "A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2002, pp. I–881–884.

[45] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. on Speech and Audio Processing*, vol. 12, no. 5, pp. 530–538, 2004.

[46] L. C. Parra and C. Spence, "On-line blind source separation of non-stationary signals," *Journal of VLSI Signal Processing*, vol. 26, no. 1/2, pp. 39–46, 2000.

[47] Y. Mori, H. Saruwatari, T. Takatani, S. Ukai, K. Shikano, T. Hiekata, and T. Morita, "Real-time implementation of two-stage blind source separation combining SIMO-ICA and binary masking," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2005, pp. 229–232.

[48] C. Avendano and J.-M. Jot, "Frequency domain techniques for stereo to multichannel upmix," in *Proc. AES 22nd Conf. on Vitual, Synthetic and Entertainment Audio*, 2002, pp. 121–130.

[49] H. Viste and G. Evangelista, "On the use of spatial cues to improve binaural source separation," in *Proc. Int. Conf. on Digital Audio Effects (DAFx)*, 2003, pp. 209–213.

[50] N. Roman and D. L. Wang, "Binaural tracking of multiple moving sources," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2003, pp. V–149–152.

[51] J. M. Peterson and S. Kadambe, "A probabilistic approach for blind source separation of underdetermined convolutive mixtures," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2003, pp. VI–581–584.

[52] S. Winter, H. Sawada, S. Araki, and S. Makino, "Overcomplete BSS for convolutive mixtures based on hierarchical clustering," in *Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA)*, 2004, pp. 652–660.

[53] T. T. Kristjánsson, J. R. Hershey, P. A. Olsen, S. J. Rennie, and R. A. Gopinath, "Super-human multi-talker speech recognition: The IBM 2006 speech separation challenge system," in *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, 2006, pp. 97–100.

[54] A. Ozerov, P. Philippe, R. Gribonval, and F. Bimbot, "One microphone singing voice separation using source-adapted models," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2005, pp. 90–93.

[55] M. A. Casey and A. Westner, "Separation of mixed audio sources by independent subspace analysis," in *Proc. Int. Computer Music Conf. (ICMC)*, 2000, pp. 154–161.

[56] M. Kim and S. Choi, "Monaural music source separation: nonnegativity, sparseness and shift-invariance," in *Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA)*, 2006, pp. 617–624.

[57] T. Virtanen, "Separation of sound sources by convolutive sparse coding," in *Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing (SAPA)*, 2004.

[58] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2003, pp. 177–180.

[59] T. Virtanen, "Sound source separation using sparse coding with temporal continuity objective," in *Proc. Int. Computer Music Conf. (ICMC)*, 2003, pp. 231–234.

[60] S. Vembu and S. Baumann, "Separation of vocals from polyphonic audio recordings," in *Proc. Int. Conf. on Music Information Retrieval (ISMIR)*, 2005, pp. 337–344.

[61] L. Benaroya, L. McDonagh, F. Bimbot, and R. Gribonval, "Non negative sparse representation for Wiener based source separation with a single sensor," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2004, pp. VI–613–616.

[62] A. S. Bregman, *Auditory Scene Analysis*. MIT Press, 1990.

[63] M. Cooke and D. P. W. Ellis, "The auditory organization of speech and other sources in listeners and computational models," *Speech Communication*, vol. 35, pp. 141–177, 2001.

[64] D. L. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. Wiley, 2006.

[65] D. K. Mellinger, "Event formation and separation in musical sound," Ph.D. dissertation, CCRMA, Stanford University, 1991.

[66] K. Kashino, K. Nakadai, T. Kinoshita, and H. Tanaka, "Application of bayesian probability network to music scene analysis," in *Working notes of Int. Joint Conf. on Artificial Intelligence (IJCAI) Workshop on Computational Auditory Scene Analysis*, 1995, pp. 52–59.

[67] D. P. W. Ellis, "Prediction-driven computational auditory scene analysis," Ph.D. dissertation, Dept. of Electrical Engineering and Computer Science, MIT, 1996.

[68] J. P. Barker, M. P. Cooke, and D. P. W. Ellis, "Decoding speech in the presence of other sources," *Speech Communication*, vol. 45, pp. 5–25, 2005.

[69] T. Kinoshita, S. Sakai, and H. Tanaka, "Musical sound source identification based on frequency component adaptation," in *Proc. Int. Joint Conf. on Artificial Intelligence (IJCAI) Workshop on Computational Auditory Scene Analysis*, 1999, pp. 18–24.

[70] J. Nix, "Localization and separation of concurrent talkers based on principles of auditory scene analysis and multi-dimensional statistical methods," Ph.D. dissertation, Dept. of Medical Physics, University of Oldenburg, 2005.

[71] M. J. Reyes-Gomez, B. Raj, and D. P. W. Ellis, "Multi-channel source separation by factorial HMMs," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2003, pp. I–664–667.

[72] R. M. Parry and I. Essa, "Estimating the spatial position of spectral components in audio," in *Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA)*, 2006, pp. 666–673.

[73] T. Nakatani, "Computational auditory scene analysis based on residue-driven architecture and its application to mixed speech recognition," Ph.D. dissertation, Dept. of Applied Analysis and Complex Dynamical Systems, Kyoto University, 2002.

[74] L. A. Drake, "Sound source separation via computational auditory scene analysis (CASA)-enhanced beamforming," Ph.D. dissertation, Dept. of Electrical Engineering, Northwestern University, 2001.

[75] Y. Sakuraba and H. G. Okuno, "Note recognition of polyphonic music by using timbre similarity and direction proximity," in *Proc. Int. Computer Music Conf. (ICMC)*, 2003, pp. 167–170.

[76] E. Vincent, "Musical source separation using time-frequency source priors," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 91–98, 2006.

[77] H. Viste and G. Evangelista, "An extension for source separation techniques avoiding beats," in *Proc. Int. Conf. on Digital Audio Effects (DAFx)*, 2002, pp. 71–75.

[78] R. V. Balan and J. P. Rosca, "Convolutive demixing with sparse discrete prior models for Markov sources," in *Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA)*, 2006, pp. 544–551.

[79] W. Herbordt and W. Kellermann, "Frequency-domain integration of acoustic echo cancellation and a generalized sidelobe canceller with improved robustness," *European Trans. on Telecommunications*, vol. 13, no. 2, pp. 123–132, 2002.

[80] S. Araki, S. Makino, Y. Hinamoto, R. Mukai, T. Nishikawa, and H. Saruwatari, "Equivalence between frequency-domain blind source separation and frequency-domain adaptive beamforming for convolutive mixtures," *EURASIP Journal on Applied Signal Processing*, vol. 11, pp. 1157–1166, 2003.

[81] J. J. Burred and T. Sikora, "On the use of auditory representations for sparsity-based sound source separation," in *Proc. IEEE Int. Conf. on Information, Communications and Signal Processing (ICICS)*, 2005, pp. 1466–1470.

[82] M. R. Every, "Separation of musical sources and structure from single-channel polyphonic recordings," Ph.D. dissertation, Dept. of Electronics, University of York, 2006.

[83] N. Roman and D. L. Wang, "Pitch-based monaural segregation of reverberant speech," *Journal of the Acoustical Society of America*, vol. 120, no. 1, pp. 458–469, 2006.