Oracle estimators for the benchmarking of source separation algorithms

Emmanuel Vincent, Rémi Gribonval, Mark Plumbley

Emmanuel Vincent [1,*] Rémi Gribonval

METISS group
IRISA-INRIA
Campus de Beaulieu, 35042 Rennes CEDEX, France

Mark D. Plumbley

Center for Digital Music, Department of Electronic Engineering
Queen Mary, University of London
Mile End Road, London E1 4NS, United Kingdom

Oracle Estimators for the Benchmarking of Source Separation Algorithms

Abstract

Source separation is a difficult problem for which many algorithms have been proposed. In this article, we define oracle estimators which compute the best performance achievable by different classes of algorithms on a given mixture, in a theoretical evaluation framework where the reference sources are available. We describe explicit oracle estimators for three particular classes of algorithms: multichannel time-invariant filtering, single-channel time-frequency masking and multichannel time-frequency masking. We evaluate their performance on various audio mixtures and study their robustness. We draw several conclusions for their three typical applications, namely providing performance bounds for existing and future blind algorithms, selecting the best class of algorithms for a given mixture and assessing the separation difficulty. In particular, we show that it is worth developing blind time-frequency masking algorithms relaxing the common assumption of a single active source per time-frequency point.

Key words: Blind source separation, performance, evaluation, benchmark

* Corresponding author.
Email addresses: emmanuel.vincent@irisa.fr (Emmanuel Vincent), remi.gribonval@irisa.fr (Rémi Gribonval), mark.plumbley@elec.qmul.ac.uk (Mark D. Plumbley).
[1] Emmanuel Vincent was funded by EPSRC grant GR/S75802/01.

1 Introduction

Most audio, video and biomedical signals are mixtures of several sources that are active simultaneously. For static point sources, the mixing process generally consists of a linear time-invariant filtering of the source signals. The ith channel of the observed mixture ($1 \leq i \leq I$) is then expressed as

$$x_i(t) = \sum_{j=1}^{J} \sum_{\tau=-\infty}^{+\infty} a_{ij}(\tau) s_j(t-\tau) \tag{1}$$

where $s_j(t)$, $1 \leq j \leq J$, are the source signals and $a_{ij}(\tau)$ the mixing filters. The mixture is termed *instantaneous* when the mixing filters are simple gains, *anechoic* when they are (possibly fractional) delay filters multiplied by some gains, and *convolutive* otherwise. It is also termed *over-determined* when the number of observed channels I is larger than the number of sources J, *determined* when it is equal and *under-determined* when it is smaller. The study of mixture signals raises the problem of source separation, that is the estimation of each source signal with the best possible quality.

Many algorithms have been proposed to solve this problem. Determined or over-determined mixtures are generally separated by *multichannel time-invariant filtering*, which rejects interference from certain spatial directions by applying linear demixing filters to the mixture channels [1]. Independent Component Analysis (ICA) estimates the demixing filters by assuming that the source signals are independent and non-Gaussian [2,1] or Gaussian with non-stationary variance [3]. Other approaches rely on more complex source models that incorporate detailed prior information about a specific source [4]. Under-determined mixtures are more often separated using *time-frequency masking* methods, such as binary masking [5] or adaptive Wiener filtering [6], which attenuate or remove interference in selected time-frequency points. The time-frequency masks are usually derived from the intensity and phase difference between the mixture channels [7,8], or estimated using a model of the short-term power spectra of the sources [5,6,9].

The performance of these various algorithms exhibits wide variations depending on the properties of the sources and the mixing filters. For example, the performance of convolutive ICA algorithms is generally high on anechoic mixtures, but decreases quickly when mixing filters become longer than a few thousand taps [10]. Three main factors can explain this experimental observation [11]:

- The *constraints* inherent to the class of separation algorithms, such as the restriction to time-invariant demixing filters and a limited filter length, may set an upper bound on the best possible performance.

- The parameters which optimize the chosen *objective function*, *i.e.* the filters which maximize the independence of the resulting sources, may not be optimal in terms of separation performance.
- The *optimization algorithm* itself may fail to maximize the objective function, perhaps due to local maxima.

In order to improve upon existing ICA algorithms, it is necessary to understand the relative importance of these factors. Quantifying this relative importance would allow research to be focused on modifying the time-invariant filtering assumption, designing improved objective functions or building better optimization algorithms, as appropriate. The same reasoning also applies to other separation algorithms than ICA.

In this article, we address this question regarding the first factor by designing algorithms to determine the separation parameters providing the best possible performance under simple constraints. Following the terminology in statistics, these algorithms are called *oracle estimators* and the resulting source estimates are called *oracle estimates*. By definition, the use of oracle estimators is restricted to an evaluation context where reference source signals are available. The study of these estimators leads to three typical applications [11]: providing theoretical upper bounds on the performance of existing and future blind algorithms, predicting the adequacy of various classes of algorithms for a given mixture signal and quantifying the difficulty of separating this signal.

Note that the influence of other factors on the separation performance is more difficult to quantify. Indeed computing the global maximum of the objective function would require prior knowledge of the number of local maxima of this function or use of a perfect optimization algorithm. For certain objective functions, additional performance bounds may be obtained using information theory [2] for example. This issue is not considered in the following. By contrast, oracle estimators are not specific to a particular objective function.

Our approach builds upon the pioneering studies [12,13,14,11,15], which focused on computing demixing filters for determined or over-determined convolutive mixtures by optimizing different criteria given reference data. Strictly speaking, the filters developed in [12,13,14,11] are not oracle estimates, since the optimized criteria differ from the chosen performance measures. The recent study [15] addresses this issue, but relies on a performance measure specific to time-invariant filtering that is not suited for other classes of algorithms. In the following, we provide a more general framework for the definition of oracle estimators and we compare the resulting oracle demixing filters to those proposed in [12,13,14]. We also design oracle estimators for two other classes of algorithms, namely single-channel and multichannel time-frequency masking. Finally, we compare all these estimators on various types of audio mixtures and we study their robustness.

The structure of the rest of the article is as follows. We start by defining oracle estimators in section 2 and explaining their use in the context of source separation. In section 3, we describe the database of example audio mixtures used in subsequent sections. Then, in sections 4, 5 and 6, we present oracle estimators for various classes of source separation algorithms and evaluate their performance separately. We provide an experimental comparison of these estimators and some usual blind separation algorithms in section 7 to illustrate the three typical applications mentioned above, and we analyze the robustness of two estimators in section 8. Finally, we summarize the contributions of this article in section 9 and point out further research directions.

2 Oracle *vs.* blind source separation

2.1 Principle

Suppose we have an observed signal $\mathbf{x}(t) = [x_1(t), \ldots, x_I(t)]^T$ from which we wish to estimate a target signal $\mathbf{y}(t) = [y_1(t), \ldots, y_C(t)]^T$, where $(\cdot)^T$ denotes transposition. Mathematically, we wish to find a separating function Φ such that the estimate $\widehat{\mathbf{y}}$ of the sequence $\mathbf{y} = [\mathbf{y}(1), \ldots, \mathbf{y}(T)]$ of length T is given by $\widehat{\mathbf{y}} = \Phi(\mathbf{x})$ where $\mathbf{x} = [\mathbf{x}(1), \ldots, \mathbf{x}(T)]$. In practice, this overly general formulation is not particularly helpful. Instead, the separation algorithms mentioned above correspond to parametric separating functions of the form

$$\widehat{\mathbf{y}} = f(\mathbf{x}, \theta) \quad \text{with } \theta \in \Theta, \tag{2}$$

where f is a fixed function, θ a vector of parameters adapted to the observed signal and Θ a set of acceptable parameters. Hence the separation problem can be broken into two steps:

- Choice of a function f and a set of constraints Θ over the parameters,
- Identification of suitable parameters θ given the observed signal $\mathbf{x}(t)$, the function f and the constraints Θ, according to some algorithm.

Each function f defines a different class of algorithms. For example, in the case of multichannel time-invariant filtering, f represents convolution, θ contains the coefficients of the demixing filters and Θ defines constraints over the length of the demixing filters or the values of particular filter coefficients.

Assuming that the target signal $\mathbf{y}(t)$ is known, the separation performance of a given algorithm can be evaluated by measuring the quality of the estimated signal using a distortion measure $d(\mathbf{y}, \widehat{\mathbf{y}})$. We define the *oracle estimator* $\widetilde{\theta}(\mathbf{y}, \mathbf{x}, \Theta)$ to be the vector of parameters resulting in the smallest distortion

among the set of acceptable parameters Θ:

$$\widetilde{\theta}(\mathbf{y}, \mathbf{x}, \Theta) = \arg\min_{\theta \in \Theta} d(\mathbf{y}, f(\mathbf{x}, \theta)). \tag{3}$$

The study of this oracle estimator consists in computing the corresponding distortion $\tilde{d}(\mathbf{y}, \mathbf{x}, \Theta)$ as a function of Θ

$$\tilde{d}(\mathbf{y}, \mathbf{x}, \Theta) = d(\mathbf{y}, f(\mathbf{x}, \widetilde{\theta}(\mathbf{y}, \mathbf{x}, \Theta))) = \min_{\theta \in \Theta} d(\mathbf{y}, f(\mathbf{x}, \theta)). \tag{4}$$

By definition, $\tilde{d}(\mathbf{y}, \mathbf{x}, \Theta)$ is the tightest possible lower bound on the distortion over all algorithms from class f.

2.2 Examples

In practice, the target signal $\mathbf{y}(t)$ may assume different forms dictated by the application. For example, it may be a single-channel signal, such as the jth source signal $s_j(t)$ or the *image* of the jth source on the ith mixture channel, defined by [16]

$$s_{ij}^{\text{img}}(t) = \sum_{\tau=-\infty}^{+\infty} a_{ij}(\tau) s_j(t-\tau), \tag{5}$$

which satisfies $x_i(t) = \sum_{j=1}^{J} s_{ij}^{\text{img}}(t)$. Alternatively, it may be a new mixture signal involving the same sources mixed differently (*remix*), or a multichannel signal consisting of all the sources $\mathbf{s}(t) = [s_1(t), \ldots, s_J(t)]^T$ or the images of all the sources on all channels $\mathbf{s}^{\text{img}}(t) = [s_{1,1}^{\text{img}}(t), \ldots, s_{1,J}^{\text{img}}(t), \ldots, s_{I,1}^{\text{img}}(t), \ldots, s_{I,J}^{\text{img}}(t)]^T$. The latter is required for applications based on 3D source estimates, such as audio wavefield analysis, or for mixtures of spatially extended sources that cannot be represented by single-channel signals. It is also a common target for blind source separation algorithms [17,18], which can estimate source images without any theoretical indeterminacy [16,17] but would suffer from scaling [2] or filtering [1] indeterminacies when estimating single-channel source signals instead. Reference source images can be acquired for synthetic mixtures by filtering the source signals with known mixing filters and for live microphone recordings by recording the sources one at a time [19,11]. Most of the derivations proposed in the following are valid for any target signal, but for the sake of consistency we choose $\mathbf{y}(t) = \mathbf{s}^{\text{img}}(t)$ in all our examples and experiments.

Note that the knowledge of the target signal suffices to define an oracle estimator. For instance, if the target signal contains a subset of the source signals, knowledge of other source signals or mixing filters is not in theory required to compute oracle separation parameters. Depending on the constraints over the parameters, the oracle parameters corresponding to different sources may be related or not. If not, the same results can be obtained using a separate oracle estimator for each source.

2.3 Distortion measure

The distortion measure $d(\mathbf{y}, \widehat{\mathbf{y}})$ between a target and its estimate is also dictated by the application [20]. In the following, we use the Euclidean distortion measure

$$d(\mathbf{y}, \widehat{\mathbf{y}}) = \|\widehat{\mathbf{y}} - \mathbf{y}\|^2 \tag{6}$$

where $\|\mathbf{a}\|^2 = \sum_{c=1}^{C} \sum_{t=0}^{T-1} a_c(t)^2$ is the squared Euclidean norm of a signal $\mathbf{a}(t)$ with C channels and T samples. We assess the overall separation performance using the Signal-to-Distortion Ratio (SDR) expressed in decibels (dB)

$$\mathrm{SDR} = 10 \log_{10} \frac{\|\mathbf{y}\|^2}{\|\widehat{\mathbf{y}} - \mathbf{y}\|^2}. \tag{7}$$

Minimizing the distortion $d(\mathbf{y}, \widehat{\mathbf{y}})$ is equivalent to maximizing the SDR. The SDR incorporates all possible kinds of distortion arising from different source separation algorithms, including interference from other sources, "gurgling" artifacts, filtering distortion and spatial distortion. It is particularly relevant for applications such as high-fidelity music remixing or source speaker identification where even spatial distortion or time-invariant filtering distortion of the target are considered disturbing.

We chose this measure in this article for two reasons: it allows a fair comparison of time-invariant filtering and time-frequency masking algorithms and it results in exact closed-form expressions for some basic oracle estimators. By contrast, the Signal-to-Interference Ratio (SIR) criterion used in some studies on time-invariant filtering algorithms [11] could provide high ratings for time-frequency masking oracles despite strong time-varying filtering distortions considered unacceptable for most applications. Alternative performance measures allowing time-invariant filtering distortions only [20] or alternative definitions of the SDR involving nonlinear averaging of squared Euclidean distances could be more relevant for other applications, but do not lead to exact closed form expressions for these estimators. Note that, when the estimated signals are audio signals destined to be listened to, the perceptual relevance of the oracle estimators could be slightly improved using a perceptually weighted Euclidean distortion measure, such as the one defined in [21].

3 Experimental data

In order to illustrate the use of the oracle estimators defined subsequently, we designed a database of simulated recordings and synthetic mixtures representing a large range of audio signals, allowing precise control of the mixing filters. The database contained equal amounts of music and speech data, avoiding un-

realistic data such as MIDI-synthesized signals or mixtures of unsynchronized solo music recordings.

We collected ten multitrack music recordings (*master tapes*) from different genres [2], each containing three sources playing synchronously and in harmony. We also selected speech sources from thirty English speakers from as many different audio books [3]. The latter were grouped into mixtures of three sources so that three mixtures contained male speakers only, three others female speakers only and the remaining four both male and female speakers. All the source signals were sampled at 22.05 kHz and truncated to 2^{18} samples (11.9 s).

Multichannel recordings of several sources were simulated by convolving the source signals with room impulse responses determined by the image technique [22] using the Roomsim toolbox [4]. The positions of the microphones and the loudspeakers are illustrated in figure 1. The number of sources was either two or three, and the number of mixture channels varied between two and eight. Mixtures with J sources and I channels involved loudspeakers numbered 1 to J and microphones numbered 1 to I only. The mixing filters were shifted by 64 samples, so that the delay corresponding to direct sound varied between -5 and +5 samples for the first two microphones. The length of the mixing filters was assessed by the room reverberation time RT, defined to be the delay after which the magnitude of the sound reflections on the room surfaces becomes 60 dB smaller than that of the original sound. This quantity is typically on the order of 250 ms in a quiet meeting room and 2 s in a concert hall. In the following, we chose RT between four values: 0 ms (anechoic), 50 ms (1100 samples at 22.05 kHz), 250 ms (5500 samples) and 1.25 s (28000 samples).

We also generated single-channel (mono) mixtures by simply adding the sources together and two-channel (stereo) instantaneous mixtures by mixing the sources with positive gains forming the matrix

$$\mathbf{A} \simeq \begin{bmatrix} 0.212 & 0.949 & 0.643 \\ 0.977 & 0.316 & 0.766 \end{bmatrix}. \tag{8}$$

All the experimental data mentioned above were made available under Creative Commons licenses as part of a toolbox called BSS Oracle at the address `http://bass-db.gforge.inria.fr/bss_oracle/`. This toolbox also contains

[2] These recordings were downloaded from the artists' websites under Creative Commons licenses. The artists are Alex Q, Another Dreamer, Brian Smith, Carl Leth, Espi Twelve, Jim's Big Ego, Mister Mouse and Mokamed. Musical genres include folk, acoustic pop, pop rock, metal, techno, electronica, trip hop and hip hop.

[3] These were chosen among public domain audio books from http://librivox.org/

[4] http://media.paisley.ac.uk/~campbell/Roomsim/

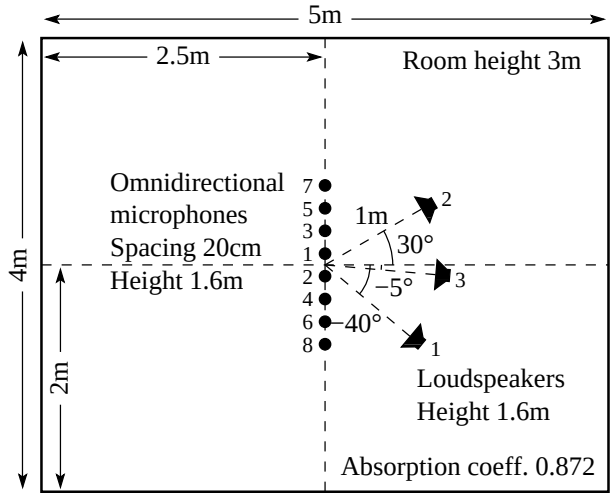

Figure 1. Microphone and loudspeaker positions for simulated room recordings with RT = 50 ms. Anechoic recordings were generated using the same configuration with an absorption coefficient of 1. Recordings with RT = 250 ms and RT = 1.25 s were simulated using the same microphone and loudspeaker positions, but with twice and four times larger rooms and absorption coefficients of 0.561 and 0.280 respectively.

Matlab programs distributed under the GNU Public License to compute oracle estimators and plot the figures of this article.

In the experiments of sections 4 to 6 we measure the average performance trends of the oracle estimators applied to this data. We provide a more detailed account for the variation of SDR depending on the data using confidence bounds in the experiments of sections 7.1 to 7.4. The experiments of sections 7.5 and 8 are based on additional data introduced in these sections.

4 Multichannel time-invariant filtering oracle

Source separation algorithms for determined or over-determined mixtures, *e.g.* [2,1,17,18,4], are generally based on time-invariant multichannel filtering, which relies mainly on the spatial diversity of the sources and to a lesser degree on their spectral diversity. Each channel of the target is estimated by filtering the mixture channels with time-invariant linear filters, called demixing filters, and summing the filtered channels together. Finite Impulse Response (FIR) demixing filters are most often used, although other filter structures may provide a competitive performance [11]. By carefully choosing these filters, it is possible to attenuate sounds coming from certain spatial positions or directions and at certain frequencies where interference dominates.

4.1 Definition of the separating function

Time-invariant filtering can be expressed either in the time domain or in the frequency domain. Let us describe the time-domain implementation first and leave the frequency-domain implementation for later consideration in section 4.4. Assuming that the demixing filters $w_{jk}(\tau)$ are non-causal filters of even length L centered at zero-lag, the estimate of source j can be written as

$$\widehat{s}_j(t) = \sum_{k=1}^{I} \sum_{\tau=-L/2+1}^{L/2} w_{jk}(\tau) x_k(t-\tau). \tag{9}$$

Similarly, the image of source j on mixture channel i may be estimated using demixing filters $w_{ijk}(\tau)$ by

$$\widehat{s}_{ij}^{\,\mathrm{img}}(t) = \sum_{k=1}^{I} \sum_{\tau=-L/2+1}^{L/2} w_{ijk}(\tau) x_k(t-\tau). \tag{10}$$

Some algorithms derive the image demixing filters $w_{ijk}(\tau)$ from the source demixing filters $w_{jk}(\tau)$ by matrix computation [17]. Other algorithms estimate unconstrained image demixing filters $w_{ijk}(\tau)$ directly using a cost function to assess the quality of the reconstructed mixture channels [18].

Extending formulations (9) and (10) to any target signal $\mathbf{y}(t)$, each channel of the target is estimated as a linear combination of the form

$$\widehat{y}_c(t) = \sum_{\eta=1}^{D} w_{c\eta} x_\eta(t) \tag{11}$$

where $\eta = (k, \tau)$ is an index varying between 1 and $D = IL$, $w_{c\eta}$ are the demixing coefficients and $x_\eta(t)$ denotes the delayed mixture channels defined by $x_\eta(t) = x_k(t-\tau)$.

4.2 Computation of the oracle parameters

The demixing coefficients $\widetilde{w}_{c\eta}$ which maximize the SDR are the solution of a separate linear least-squares problem for each channel c of the target. Classically, this solution is given by the coefficients of the orthogonal projection of the target y_c onto the subspace spanned by the delayed mixture channels x_η [23]. More explicitly, denoting by $\langle a, b \rangle = \sum_{t=0}^{T-1} a(t)b(t)$ the Euclidean inner product of two real single-channel signals a and b of length T, the vector of oracle coefficients $\widetilde{\mathbf{w}}_c = [\widetilde{w}_{c,1}, \ldots, \widetilde{w}_{c,D}]^T$ is equal to

$$\widetilde{\mathbf{w}}_c = \mathbf{G}^{-1} \mathbf{r}_c \tag{12}$$

where $\mathbf{G}$ and $\mathbf{r}_c$ are respectively the Gram matrix of the delayed mixture channels and the vector of their inner products with the target defined by $G_{\eta\eta'} = \langle x_\eta, x_{\eta'} \rangle$ and $r_{c\eta} = \langle y_c, x_\eta \rangle$ with $1 \leq \eta \leq D$, $1 \leq \eta' \leq D$.

Note that the "diagonal" demixing filters $w_{iji}(\tau)$ are not constrained to be Dirac delta functions and that their oracle estimates do not generally satisfy this constraint. Nevertheless, a similar solution could be derived for constrained filters or alternative filter structures as in [11].

Most previous studies on the performance of time-invariant filtering algorithms relied on near-optimal source demixing filters computed by pseudo-inversion of the mixing filter system [12,14], which can easily be shown to be identical to the oracle demixing filters for uncorrelated white noise sources. A recent study derived near-optimal source demixing filters by minimizing the energy of interference [11], taking in account the power spectral densities of interference sources to cancel them in priority at the frequencies where they have more energy. By contrast, the proposed oracle estimator takes into account the power spectral densities of all sources and always achieves a better SDR by allowing in addition a larger relative distortion of the target source at the frequencies where it has little energy. Using the data of section 3 and a filter length of 2048, oracle demixing filters improve the SDR by an average 1.6 dB for speech sources and 6.1 dB for music sources compared to demixing filters computed by pseudo-inversion of the mixing system.

4.3 Example applications: effect of reverberation time and over-determinacy

The proposed oracle estimator helps to quantify the factors influencing the performance of separation algorithms based on time-invariant filtering. We evaluated the performance of oracle demixing filters on two-source mixtures for various reverberation times RT and numbers of channels I.

Figure 2 shows that time-invariant filtering can potentially provide a SDR of 20 dB or more for determined mixtures with short reverberation time up to RT = 50 ms (1100 samples) using demixing filters of a few hundred taps. However its performance deteriorates for realistic reverberation times, where oracle demixing filters of a few hundred taps result in a SDR of 15 to 20 dB only. In this context, the SDR does not increase much by increasing the length of the demixing filters. This can be explained by the fact that time-invariant filtering can reject interference only from at most one direction at low frequencies and in a two-channel mixture [10]. Thus virtual interference sources generated by reverberation at random spatial directions cannot be perfectly cancelled whatever filter length is used.

When the number of mixture channels is increased, demixing filters can have

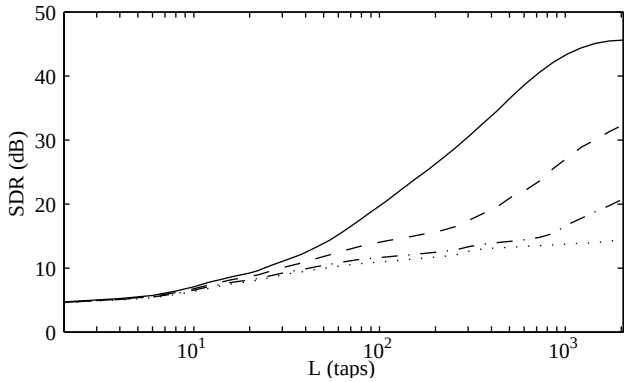

Figure 2. Average performance of the time-invariant filtering oracle on determined two-source mixtures as a function of the length L of the demixing filters. Each curve corresponds to a different reverberation time (plain: anechoic, dashed: RT = 50 ms, dash-dotted: RT = 250 ms, dotted: RT = 1.25 s).

more complex spatial responses and reject interference from several unrelated positions at each frequency. Figure 3 shows that the performance of oracle filters on convolutive two-source mixtures improves monotonically as a function of the number of mixture channels. This has already been observed using near-optimal demixing filters in [12,13].

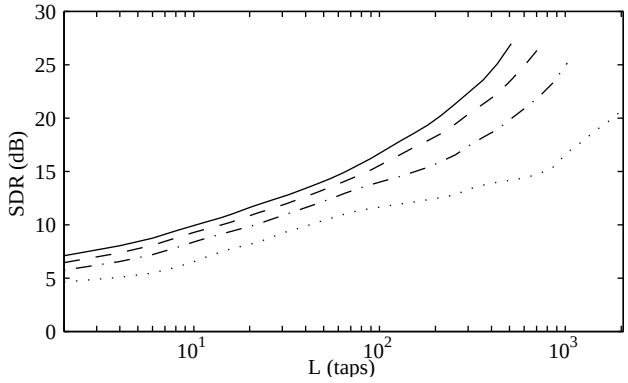

Figure 3. Average performance of the time-invariant filtering oracle on over-determined two-source mixtures with reverberation time RT = 250 ms as a function of the length L of the demixing filters. Each curve corresponds to a different number of mixture channels (plain: $I = 8$, dashed: $I = 6$, dash-dotted: $I = 4$, dotted: $I = 2$). Oracle filters with a large number of coefficients per target LI could not be estimated due to large memory requirements.

4.4 Extension to frequency-domain implementation

For computational convenience, source separation algorithms often implement time-invariant filtering in the frequency domain, where convolution translates into simple complex multiplication in each frequency bin f [1]. The Short-Term

Fourier Transforms (STFTs) $S_j(n,f)$ and $S_{ij}^{\mathrm{img}}(n,f)$ of the sources and the source images respectively are then estimated by $\widehat{S}_j(n,f) = \sum_{k=1}^{I} w_{jk}(f) X_k(n,f)$ and $\widehat{S}_{ij}^{\mathrm{img}}(n,f) = \sum_{k=1}^{I} w_{ijk}(f) X_k(n,f)$ in each time frame n, where $X_i(n,f)$ are the STFTs of the mixture channels and $w_{jk}(f)$ and $w_{ijk}(f)$ binwise complex demixing coefficients. The estimated source waveforms are derived by inverse STFT using the overlap-add technique [24]. As explained in [25], this formulation is not exactly equivalent to time-domain filtering, even when the number of frequency bins equals the length of the time-domain filters, since linear convolution is replaced by circular convolution.

The subsequent loss of performance could be evaluated by computing oracle demixing coefficients, which amounts to solving a large linear least squares problem, as in the time domain. Instead, we computed near-optimal demixing coefficients by minimizing the distortion between the STFT coefficients of the target and its estimate in each frequency bin separately. This follows the principle of many blind algorithms, which estimate the coefficients by optimizing a separate objective function for each frequency bin [1]. Thus it is expected that these near-optimal coefficients provide an upper bound on the performance of many blind algorithms.

Figure 4 shows that the time-domain oracle performs 2 dB better than its frequency-domain counterpart on average for determined two-source mixtures, or alternatively frequency-domain filters must contain about twice as many coefficients as time-domain filters to achieve a similar performance. This non-negligible difference justifies the investigation of exact formulations of time-invariant filtering in the frequency domain, such as the one proposed in [25].

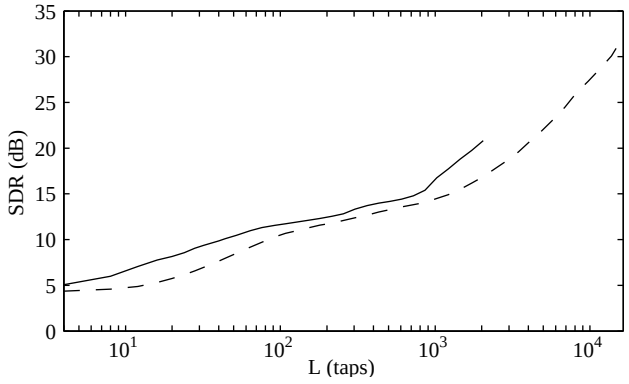

Figure 4. Average performance of various source image demixing filters on determined two-source mixtures with reverberation time RT = 250 ms. Plain: time-domain oracle demixing filters of length L. Dashed: near-optimal frequency-domain demixing filters with L frequency bins. Long oracle filters could not be estimated due to large memory requirements.

5 Single-channel time-frequency masking oracle

Time-invariant filtering requires at least as many mixture channels as there are sources, so it is not suited for under-determined mixtures, and in particular for single-channel mixtures. Source separation algorithms for single-channel mixtures, *e.g.* [5,6,9], are usually based on time-frequency masking, which is a particular kind of non-stationary filtering conducted in the time-frequency domain and relying on the time-frequency diversity of the sources. By carefully designing the time-varying magnitude response of the filters, it is possible to filter out time-frequency regions dominated by interference.

5.1 Definition of the separating function

Masking can be conducted on any time-frequency representation of the data, including the widely used STFT. For simplicity, we postpone consideration of the STFT until section 5.4 and assume instead that time-frequency masking is performed using an orthonormal time-frequency basis, such as the Modified Discrete Cosine Transform (MDCT) [26]. The MDCT coefficients of the single-channel mixture signal $x(t)$ are given by $\langle x, \phi_m \rangle$, where $\phi_m(t)$, $1 \leq m \leq M$, are the elements of the MDCT basis. Each source image is estimated by multiplying these coefficients by masking coefficients ϵ_{jm} and inverting the MDCT representation by weighted summation of the basis elements. In the end, the source image estimates $\widehat{s}_j^{\text{img}}(t)$ can be written as

$$\widehat{s}_j^{\text{img}}(t) = \sum_{m=1}^{M} \epsilon_{jm} \langle x, \phi_m \rangle \phi_m(t). \tag{13}$$

Two types of masks are encountered in the literature: *binary masks* containing discrete values $\epsilon_{jm} \in \{0, 1\}$ [5] and *real-valued masks* containing gains $0 \leq \epsilon_{jm} \leq 1$ [6]. In both cases, the masking coefficients generally satisfy the unitary sum constraint

$$\sum_{j=1}^{J} \epsilon_{jm} = 1 \quad \forall m. \tag{14}$$

5.2 Computation of the oracle parameters

Extending formulation (13) to any target signal $\mathbf{y}(t)$, each channel of the target can be expressed in the MDCT basis as $y_c(t) = \sum_{m=1}^{M} \langle y_c, \phi_m \rangle \phi_m(t)$ and its estimate as $\widehat{y}_c(t) = \sum_{m=1}^{M} \epsilon_{cm} \langle x, \phi_m \rangle \phi_m(t)$. Since MDCT elements are orthonormal, the total Euclidean distortion over the target can be decomposed

as $\|\widehat{\mathbf{y}}-\mathbf{y}\|^2 = \sum_{m=1}^{M}(\epsilon_{cm}\langle x,\phi_m\rangle - \langle y_c,\phi_m\rangle)^2$, yielding after simple computation

$$\|\widehat{\mathbf{y}}-\mathbf{y}\|^2 = \sum_{m=1}^{M}\langle x,\phi_m\rangle^2\left(\sum_{c=1}^{C}(\epsilon_{cm}-r_{cm})^2\right), \tag{15}$$

where $r_{cm} = \langle y_c,\phi_m\rangle/\langle x,\phi_m\rangle$ denotes the ratio of the MDCT coefficients of the target and the mixture signal. Minimizing the total distortion is thus equivalent to minimizing the distortion over each MDCT element separately. Note that zero distortion can be achieved by setting $\epsilon_{cm} = r_{cm}$ only when $r_{cm} \in \{0,1\}$ for binary masking or when $0 \leq r_{cm} \leq 1$ for real-valued masking.

Given the unitary sum constraint (14), the oracle binary masking coefficients are found by

$$\widetilde{\epsilon}_{cm} = \begin{cases} 1 & \text{if } c = \arg\max_{c'} r_{c'm}, \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

The computation of the oracle real-valued masking coefficients is a linear least squares problem with bound and linear equality constraints. The solution can be found by combinatorial search over the faces of the constraint polytope and resolution of separate unconstrained linear least squares problems [23]. When the number of sources is large, it can also be obtained more efficiently using iterative active set or interior point methods [23].

5.3 Example application: choice of the MDCT length

Oracle estimators provide a natural framework to choose some parameters of time-frequency masking algorithms, including the length of MDCT elements. We compared the various masking oracles on single-channel two-source mixtures using MDCT bases built from sine windows [26]. Figure 5 shows that the optimal MDCT length equals about 1200 samples (55 ms) for speech mixtures and 4100 samples (190 ms) for music mixtures. This is likely to be because music is more "stationary" than speech. Half-length or double-length MDCT results in an average SDR degradation of 0.5 dB. A similar result for speech data was obtained previously in [7] in the particular case of binary masking.

In this experiment, the maximal performance level is very similar for speech and music mixtures, with SDRs of 17.1 dB and 17.7 dB respectively. Thus the fact that the considered music sources exhibit more time-frequency overlap because they play in harmony does not prevent them from being separated by time-frequency masking as well as speech sources, even if the determination of the optimal masks may be more difficult in a blind context. Note also that real-valued masking performs 3 dB better than binary masking on average, which justifies the use of real-valued masking both for speech and music data.

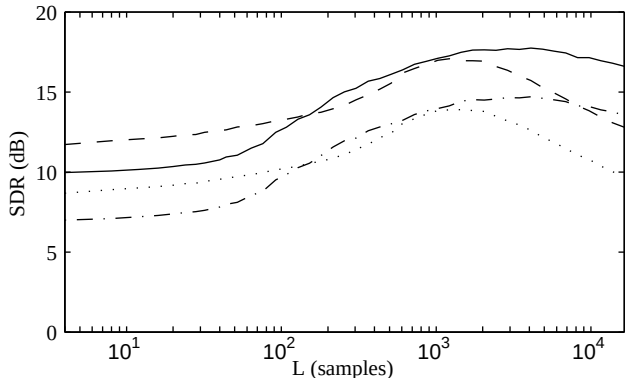

Figure 5. Average performance of the time-frequency masking oracles on single-channel two-source mixtures as a function of the length L of the MDCT basis elements. Plain: real-valued masks applied to music data. Dashed: real-valued masks applied to speech data. Dash-dotted: binary masks applied to music data. Dotted: binary masks applied to speech data.

5.4 Extension to overcomplete time-frequency transforms

In practice, time-frequency masking is often conducted using a STFT rather than the MDCT. A STFT is typically an overcomplete transform, where the over-completeness factor depends on the amount of overlap between successive time frames, assuming no zero-padding of the FFT. For instance, when standard half-overlapping windows are employed, there are twice as many STFT coefficients as samples in the time-domain signal. Denoting by $X(n,f)$ the STFT of the single-channel mixture signal, the STFTs $S_j^{\text{img}}(n,f)$ of the source images are then expressed as $\hat{S}_j^{\text{img}}(n,f) = \epsilon_j(n,f)X(n,f)$, where $\epsilon_j(n,f)$ are the masking coefficients, and the source image waveforms are recovered by STFT inversion using the overlap-add technique [24].

Due to the non-orthogonality of the STFT, oracle masking coefficients must be determined jointly in all time-frequency points using full combinatorial search for binary masks or a gradient technique for real-valued masks. Clearly, this is infeasible for realistic signals involving hundreds of thousands of samples. Instead, we obtained near-optimal masks by minimizing the distortion on the target estimate in each time-frequency point separately. This distortion takes the same form as previously, with $\|\widehat{\mathbf{Y}}(n,f)-\mathbf{Y}(n,f)\|^2 = |X(n,f)|^2 \sum_{c=1}^{C}(\epsilon_c(n,f)-R_c(n,f))^2$ up to an additive constant, where $R_c(n,f) = \Re(Y_c(n,f)/X(n,f))$ is the real part of the ratio of STFT coefficients of the target and the mixture. The masking coefficients minimizing this distortion can thus be found using the algorithms above. Note that these coefficients are different from the coefficients derived by adaptive Wiener filtering from the magnitude of the target STFTs. The latter are optimal in a probabilistic sense under the hypothesis that the targets are zero-mean Gaussian [6], but they are not guaranteed to result in the smallest possible distortion.

Figure 6 shows that the average SDR obtained with near-optimal binary masks on single-channel two-source mixtures increases by 0.6 dB when the over-completeness factor varies from 2 to 8. Similarly, the average performance increase obtained with near-optimal real-valued masks was less than 0.2 dB. This is much less than the increase reported in [27] using a blind algorithm to compute the masks. This suggests that over-completeness improves performance mainly by helping finding better masks in a blind context, rather than improving the potential performance of masking itself.

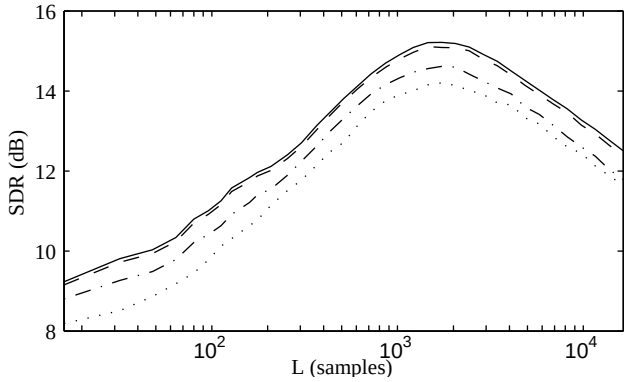

Figure 6. Average performance of the binary MDCT time-frequency masking oracle and the near-optimal STFT time-frequency masks on single-channel two-source mixtures as a function of the length L of the MDCT/STFT elements. Each curve corresponds to a different over-completeness factor (plain: eight-times overcomplete STFT, dashed: four-times overcomplete STFT, dash-dotted: twice overcomplete STFT, dotted: MDCT).

6 Multichannel time-frequency masking oracle

Besides its use for single-channel mixtures, time-frequency masking is also a common approach to multichannel source separation, as popularized by the Degenerate Unmixing Estimation Technique (DUET) [7]. Indeed it can combine the advantages of multichannel filtering and time-frequency masking by jointly exploiting spatial diversity and time-frequency diversity.

6.1 Definition of the separating functions

Multichannel time-frequency masking is most often performed on a STFT representation of the signal. However, as in the previous section, we first consider masking on an orthonormal time-frequency basis such as the MDCT and we describe the STFT implementation later in section 6.4. In the following, we focus on two different separating functions: binary masking and local mixing

inversion [28,7]. For the sake of conciseness, alternative separating functions, such as those stemming from the chaining of ICA and binary masking [29], are not considered in this article.

6.1.1 Binary masking

Multichannel binary masking is a straightforward generalization of single-channel binary masking (13) where a single mask $\epsilon_{jm} \in \{0,1\}$ is applied to all channels $x_i(t)$ of the mixture for each source j. The masks are subject to the unitary sum constraint (14). Each source image is then estimated as

$$\widehat{s}_{ij}^{\text{img}}(t) = \sum_{m=1}^{M} \epsilon_{jm} \langle x_i, \phi_m \rangle \phi_m(t). \tag{17}$$

6.1.2 Local mixing inversion

Local mixing inversion is a more advanced method that exploits all mixture channels together. Assuming that the observed signal $\mathbf{x}(t)$ is an instantaneous mixture with known mixing gains a_{ij} forming a $I \times J$ matrix $\mathbf{A}$, the coefficients of the mixture in the MDCT basis satisfy $[\langle x_1, \phi_m \rangle, \ldots, \langle x_I, \phi_m \rangle]^T = \mathbf{A}[\langle s_1, \phi_m \rangle, \ldots, \langle s_J, \phi_m \rangle]^T$. Denoting by $\mathcal{J}_m$ the set of size J'_m containing the indexes of the sources contributing most actively to the mixture at the time-frequency point m, the coefficients of the sources in the basis are estimated as [28,30]

$$\begin{cases} \widehat{\langle s_j, \phi_m \rangle} & = 0 \qquad\qquad\qquad\qquad j \notin \mathcal{J}_m, \\ \left[\widehat{\langle s_j, \phi_m \rangle}\right]_{j \in \mathcal{J}_m}^T & = \mathbf{A}_{\mathcal{J}_m}^{\dagger} \; [\langle x_i, \phi_m \rangle]_{1 \le i \le I}^T \end{cases} \tag{18}$$

where $\mathbf{A}_{\mathcal{J}_m}$ denotes the $I \times J'_m$ matrix composed of the columns $\mathbf{A}_j$ of $\mathbf{A}$ indexed by $j \in \mathcal{J}_m$, and $\mathbf{A}_{\mathcal{J}_m}^{\dagger}$ denotes its $J'_m \times I$ pseudo-inverse. The index set $\mathcal{J}_m$ is called an *activity pattern*. When only one source is assumed active, $\mathcal{J}_m$ is reduced to a single index $\{j_m\}$ and the pseudo-inverse equals $\mathbf{A}_{j_m}^T / \|\mathbf{A}_{j_m}\|_2^2$, resulting in a simple expression for the estimated source coefficients

$$\widehat{\langle s_{j_m}, \phi_m \rangle} = \frac{\mathbf{A}_{j_m}^T \; [\langle x_i, \phi_m \rangle]_{1 \le i \le I}^T}{\|\mathbf{A}_{j_m}\|_2^2}. \tag{19}$$

The estimated sources and their images on the mixture channels are eventually built from their basis coefficients as

$$\widehat{s}_j(t) = \sum_{m=1}^{M} \widehat{\langle s_j, \phi_m \rangle} \phi_m(t), \tag{20}$$

$$\widehat{s}_{ij}^{\text{img}}(t) = \sum_{m=1}^{M} a_{ij} \widehat{\langle s_j, \phi_m \rangle} \phi_m(t). \tag{21}$$

Note that, when $J'_m < I$, the observed mixture signal may be different from the sum of the estimated source images.

In practice, the difficulty of local mixing inversion lies in the blind computation of the activity patterns $\mathcal{J}_m$. While DUET [7] assumes exactly $J'_m = 1$ active source in each time-frequency point, other approaches [31,30] rely on the looser assumption that $J'_m \leq I$, *i.e.* the number of simultaneously active sources does not exceed the number of mixture channels. Allowing a free number of active sources $J'_m \leq J$ may improve performance, but makes the blind estimation of activity patterns very challenging.

6.2 Computation of the oracle parameters

Since MDCT elements $\phi_m(t)$ are mutually orthogonal, the total Euclidean distortion between any target signal and its estimate can be decomposed as a sum over each coordinate. Therefore, minimizing the total distortion amounts to optimizing the masking coefficients ϵ_{jm} or the activity patterns $\mathcal{J}_m$ for each m separately. When the target signal consists of the images of all the sources on all channels $\mathbf{s}^{\text{img}}(t)$, this leads to the optimization problem

$$\widetilde{\mathcal{J}_m} = \arg\min_{\mathcal{J}_m \in \mathcal{P}} \sum_{j=1}^{J} \sum_{i=1}^{I} \left(\langle \widehat{s}_{ij}^{\text{img}}, \phi_m \rangle - \langle s_{ij}^{\text{img}}, \phi_m \rangle \right)^2 \tag{22}$$

where the role of the activity pattern $\mathcal{J}_m$ in the right hand side is implicit (see (18), (20), (21)). The set of allowed activity patterns $\mathcal{P}$ can typically be the set of all activity patterns with exactly or at most J' active sources. This is a combinatorial problem which can be addressed by exhaustive search over all possible activity patterns when J' is small.

Note that, if the assumption of a known mixing matrix $\mathbf{A}$ were relaxed, it would still be possible to define a joint estimator of the mixing matrix and the activity patterns resulting in the best performance. However, its exact computation would involve joint combinatorial optimization of the activity patterns at all time-frequency points. Since this is infeasible for realistic signals, we maintain the assumption of known $\mathbf{A}$ in the following.

6.3 Example application: choice of the number of active sources

Oracle estimators provide some insight into the role of the MDCT length and the assumed number of active sources on the performance achievable by multichannel time-frequency masking. We evaluated the performance of oracle local

mixing inversion on two-channel three-source instantaneous mixtures considering three cases: a free number J'_m of active sources in each time-frequency point m, exactly $J' = 2$ active sources or exactly $J' = 1$ active sources. For comparison purposes, we also computed the performance of oracle binary masking. Figure 7 shows that allowing two active sources per time-frequency point instead of one can improve the SDR by 10 dB with the optimal MDCT length. Allowing a free number of active sources improves the SDR by an additional 1.5 dB only. The optimal MDCT length equals about 1200 samples (55 ms) whatever the number of active sources, and performance varies in a similar way to that in the single-channel experiments illustrated in figure 5 when the length is increased or decreased.

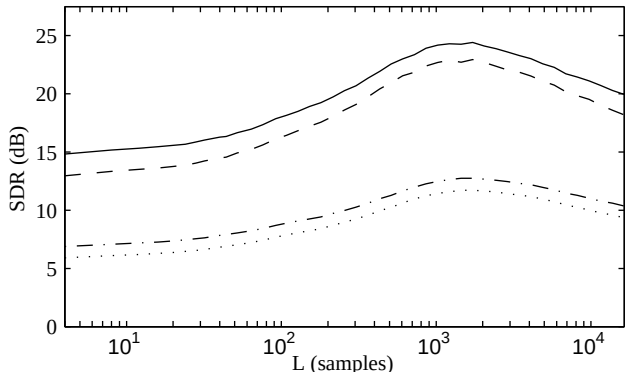

Figure 7. Average performance of the multichannel time-frequency masking oracles on two-channel three-source instantaneous mixtures as a function of MDCT length L. Plain: local mixing inversion with a free number of active sources J'_m for each time-frequency point m. Dashed: $J' = 2$ active sources. Dash-dotted: $J' = 1$ active source. Dotted: binary masking.

6.4 Extension to convolutive mixtures

In practice, multichannel time-frequency masking is often performed on a STFT rather than a MDCT. Binary masking is then conducted using the equation $\widehat{S}_{ij}^{\text{img}}(n,f) = \epsilon_j(n,f)X_i(n,f)$. The use of STFT allows to extend local mixing inversion to anechoic [7] and convolutive [32,33] mixtures by approximating convolution as a set of complex multiplications. More precisely, the multichannel STFTs of the mixture and source signals are related by $\mathbf{X}(n,f) \approx \mathbf{A}(f)\mathbf{S}(n,f)$, where $\mathbf{A}(f)$ is the $I \times J$ mixing matrix containing the Fourier transform coefficients of the mixing filters $a_{ij}(\tau)$ in frequency bin f. Denoting by $\mathcal{J}(n,f)$ the source activity pattern at time-frequency point (n,f), the source STFTs are estimated by local mixing inversion as

$$\begin{cases} \widehat{S_j}(n,f) & = 0 & j \notin \mathcal{J}(n,f), \\ \left[\widehat{S_j}(n,f)\right]_{j\in\mathcal{J}(n,f)} & = \mathbf{A}^{\dagger}_{\mathcal{J}(n,f)}(f)\,\mathbf{X}(n,f) & \end{cases} \tag{23}$$

and the source image STFTs are derived as $\widehat{S}_{ij}^{\mathrm{img}}(n,f) = a_{ij}(f)\widehat{S}_j(n,f)$. Source waveforms are recovered by STFT inversion using the overlap-add method [24].

Again, due to the non-orthogonality of the STFT, oracle binary masks or activity patterns must be computed jointly in all time-frequency points using a full combinatorial search. Since this is infeasible for long signals, we obtained near-optimal binary masks and activity patterns instead by separate minimization of $\|\widehat{\mathbf{S}}^{\mathrm{img}}(n,f) - \mathbf{S}^{\mathrm{img}}(n,f)\|^2$ in each time-frequency point.

Figure 8 illustrates the corresponding performance on two-channel three-source mixtures with reverberation time RT = 250 ms for various assumed numbers of active sources per time-frequency point. Similarly as for the instantaneous case, selecting $J' = 2$ active sources instead of $J' = 1$ for local mixing inversion improves SDR by 7 dB with the optimal window length, and free selection of $J'(n,f)$ for each time-frequency point improves SDR by a further 1.0 dB. However, in contrast with the instantaneous case, the optimal window length depends on the assumed number of active sources. The optimal length equals 2500 samples (110 ms) with $J' = 1$ and 7800 samples (350 ms) with $J' = 2$. This is larger than the optimal length for binary masking, which equals 1700 samples (75 ms). It is possible that the free choice of the number of active sources $J'(n,f)$, which is bound to perform consistently better than each specific choice $J' = 1$ or $J' = 2$, switches from a dominant choice $J'(n,f) = 1$ for small windows to $J'(n,f) = 2$ for large windows.

Another striking observation is that the oracle SDR with $J' = 2$ is extremely poor with short windows and becomes much smaller than the oracle SDR with $J' = 1$ below a window length of about 1200 samples (55 ms). This is related to the fact that the approximate modeling of the mixing process by $\mathbf{X}(n,f) \approx \mathbf{A}(f)\mathbf{S}(n,f)$ is more accurate for relatively large windows compared to the reverberation time. For short windows, the oracle estimation error remains small when $J' = 1$ because $\mathbf{A}_{\mathcal{J}(n,f)}(f)$ is always well-conditioned, but it can become very large when $J' = 2$ in the frequency bins f where $\mathbf{A}_{\mathcal{J}(n,f)}(f)$ is ill-conditioned for all allowed activity patterns $\mathcal{J}(n,f)$. This often happens in several frequency bins in practice. For example, if the mixing filters are simple delays τ_{ij}, the cofficients of $\mathbf{A}(f)$ are equal to $a_{ij}(f) = \exp(-2i\pi f\tau_{ij}/L)$ and $\mathbf{A}_{\mathcal{J}(n,f)}(f)$ is ill-conditioned for all the frequency bins f close to integer multiples of $L/(\tau_{1,j_1} + \tau_{2,j_2} - \tau_{1,j_2} - \tau_{2,j_1})$ with $\mathcal{J}(n,f) = \{j_1, j_2\}$. This may explain the performance degradation observed with $J' = 2$ in a previous study [32], which used short windows of 16 ms for RT = 130 ms.

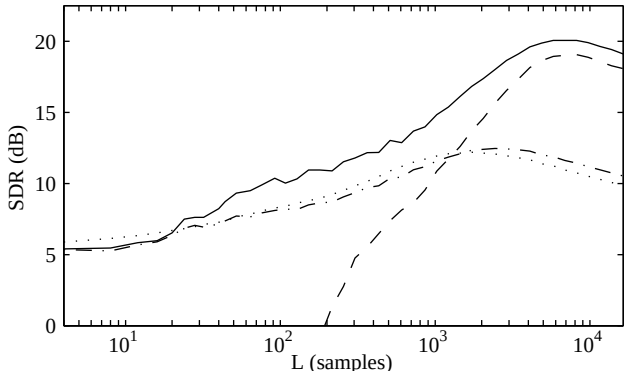

Figure 8. Average performance of multichannel near-optimal time-frequency masks on two-channel three-source mixtures with reverberation time RT = 250 ms as a function of STFT length L. Plain: local mixing inversion with a free number of active sources $J'(n, f)$ for each time-frequency point (n, f). Dashed: $J' = 2$ active sources. Dash-dotted: $J' = 1$ active source. Dotted: binary masking. The SDR range is restricted to positive values for legibility, although the dashed curve spans a larger range bounded by -30 dB.

7 Experimental comparison and typical applications

After having studied each oracle estimator separately in the previous sections, we now provide a comparison of several oracle estimators and some blind separation algorithms. Three typical applications of the oracle framework are considered [11]. Firstly, oracle performance bounds allow us to determine by how much existing blind algorithms could potentially be improved by modifying their underlying objective functions or optimization algorithms. Secondly, they indicate the potential performance of a class of separation algorithms for a given mixture, which may lead to the design of better future blind algorithms by choosing them from the appropriate class. Finally, they can provide an objective measure of the intrinsic difficulty of separating a given mixture. In the following, we discuss these applications on fourty simulated speech and music recordings and more briefly on three live microphone music recordings.

7.1 Experimental setup for simulated recordings

We selected four categories of simulated recordings among the data of section 3, each characterized by a single mixing system and a type of source: over-determined reverberant speech mixtures ($I = 4$, $J = 2$, RT = 250 ms), determined reverberant speech mixtures ($I = J = 2$, RT = 250 ms), under-determined anechoic music mixtures ($I = 2$, $J = 3$) and under-determined reverberant music mixtures ($I = 2$, $J = 3$, RT = 250 ms). Ten mixtures were obtained for each category by applying the corresponding mixing system to ten different sets of sources.

We compared near-optimal estimators for frequency-domain time-invariant filtering (see section 4.4) and multichannel binary masking (see section 6.4) with two blind source separation algorithms belonging to the same two classes: the Frequency-Domain ICA (FDICA) algorithm presented in [17] and the DUET algorithm described in [7]. We also evaluated the performance of the near-optimal estimator for local mixing inversion with a free number of active sources per time-frequency point (see section 6.4).

The original FDICA algorithm was extended to over-determined mixtures by applying principal component analysis in each frequency bin prior to ICA. Also the original DUET algorithm was modified to use binary masking instead of local mixing inversion, since we found that it provided a better performance experimentally due to large errors in the estimated mixing matrices. Note that FDICA remains limited to determined and over-determined mixtures ($J \leq I$) and DUET to stereo mixtures ($I = 2$), but that near-optimal estimators can be applied for all categories. For FDICA, near-optimal time-invariant filtering and near-optimal local mixing inversion, the STFT length was set to 4096 since this provided the best results experimentally for FDICA. The STFT length for DUET and near-optimal binary masking was set to 2048, following the optimal length for binary masking determined in section 6.4. Half-overlapping sine windows were used.

The results are shown in table 1. Each figure indicates the performance of a given algorithm or estimator averaged over the mixtures of a particular category, with 95% confidence bounds quantifying performance variability depending on the sources assuming a Gaussian performance distribution (mixing filters being fixed for each category). We discuss the main observations below.

7.2 Performance bounds of existing and future blind algorithms

The results in table 1 show that each considered blind separation algorithm performs substantially worse than the corresponding near-optimal estimator. The SDR obtained with FDICA is at least 7 dB below that of near-optimal time-invariant filtering and the SDR achieved by DUET is between 2 and 6 dB below that of near-optimal binary masking. Interestingly though, performance varies differently for blind algorithms and near-optimal estimators depending on the category of mixtures. While the performance of near-optimal time-invariant filtering increases with the number of mixture channels, that of FDICA remains almost identical, indicating that FDICA does not fully benefit from additional channels. Also, DUET performs worse on reverberant mixtures than on anechoic ones, while the near-optimal binary masking estimator performs similarly in both cases. This indicates that the reduced performance of DUET on reverberant mixtures is likely to be due to the as-

Table 1
Comparison of oracle estimators and blind source separation algorithms on simulated speech and music recordings. Bounds indicate 95% confidence intervals. FDICA is not applicable when $J > I$ and DUET when $I \neq 2$.

SDR (dB)	Speech $I = 4$, $J = 2$ RT = 250 ms	Speech $I = J = 2$ RT = 250 ms	Music $I = 2$, $J = 3$ anechoic	Music $I = 2$, $J = 3$ RT = 250 ms
FDICA	11.5 ± 1.8	11.4 ± 1.8	N/A	N/A
Near-opt. time-invar. filtering	28.3 ± 0.8	19.0 ± 0.7	16.2 ± 3.8	14.8 ± 3.1
DUET	N/A	9.3 ± 1.5	10.3 ± 2.4	7.4 ± 1.8
Near-optimal binary masking	13.9 ± 1.2	14.1 ± 1.2	12.9 ± 2.2	12.6 ± 2.3
Near-opt. local mixing inversion	22.8 ± 0.7	23.2 ± 0.8	24.8 ± 4.2	21.0 ± 3.0

sumption of anechoic mixing in the blind mask estimation stage, rather than any inability of binary masking itself.

These results suggest that it would be feasible to design improved blind time-invariant filtering or binary masking algorithms, if it were possible to find new objective functions and/or optimization algorithms approaching the oracle performance bounds. For a given objective function, it may be possible to explore how close the performance at the global optimum of that objective function approaches these bounds, at least for small problems where the parameter space could be thoroughly explored. This may also indicate how well current optimization algorithms find the global optimum of this objective function, and if necessary whether any improved optimization algorithms could be constructed.

7.3 Best class of algorithms

Comparing the performance of the various near-optimal estimators, one can observe that near-optimal time-invariant filtering exhibits better performance (by at least 2 dB) than near-optimal binary masking, even for under-determined mixtures. This is likely to be because time-invariant filtering effectively combines all mixture channels in the separation function. Local mixing inversion performs consistently better (by about 10 dB) than binary masking. This indicates that the binary masking assumption, that sources have disjoint time-frequency supports, has limited validity.

Overall, time-invariant filtering is potentially the best class of separation algorithms for over-determined speech mixtures, where it outperforms local mixing inversion by about 6 dB . For all other categories of mixtures, local mixing inversion is preferable and its performance exceeds that of time-invariant filtering by 4 to 9 dB. This suggests that it is worth developing blind local mixing inversion algorithms relaxing the assumption of a single active source per time-frequency point, despite the increased difficulty of blindly estimating the source activity patterns. So far, only a few such algorithms allowing as many active sources as mixture channels in each time-frequency point have been developed for under-determined instantaneous [31] and convolutive mixtures [33], based on l_1 norm minimization. However these algorithms rely on precise estimation of the mixing matrices, which remains a challenging problem in under-determined reverberant conditions.

7.4 Towards separation difficulty measures

Intuitively, some source separation problems are more difficult to solve than others. For example, it is generally assumed that reverberant and/or under-determined mixtures are more difficult to separate than anechoic and/or determined ones. Difficulty measures have been proposed previously, *e.g.* [34,19], but these are often specific to a particular objective function and not easily related to the achievable performance. For example, the measures in [34] are specific to ICA algorithms, since they typically measure the independence of the sources. The observed performance of oracle estimators provides a more general characterization of the actual difficulty of separating a mixture, since it gives a numerical upper bound on the performance achievable with a given class of algorithms and does not depend on a specific objective function or optimization algorithm within this class.

For example, the results in table 1 confirm that reverberant under-determined music mixtures are more difficult to separate than anechoic ones, since the performance of all oracle estimators is higher on anechoic mixtures. Interestingly though, the increase in difficulty due to reverberation appears larger for local mixing inversion than for binary masking. If separation algorithms were to be freely chosen among all classes, a single difficulty rating could be defined by selecting the maximal oracle performance among all estimators.

7.5 Application to live microphone recordings

In order to illustrate the use of oracle estimators for live microphone recordings, we repeated the experiment of section 7.1 for three music recordings made in the Espro room at IRCAM with a reverberation time of RT $\simeq$ 800 ms.

Table 2
Comparison of oracle estimators and blind source separation algorithms on live music recordings. FDICA is not applicable when $J > I$ and DUET when $I \neq 2$.

SDR (dB)	$I = 4$, $J = 2$	$I = J = 2$	$I = 2$, $J = 3$
FDICA	1.7	1.4	N/A
Near-opt. time-invar. filtering	32.9	22.4	14.4
DUET	N/A	3.4	1.8
Near-optimal binary masking	17.1	18.3	11.4
Near-opt. local mixing inversion	20.1	20.2	15.9

The reference source image signals were obtained by playing each track of a commercial multitrack music recording successively on a different loudspeaker and recording it synchronously on four microphones. The loudspeakers were placed on a circle with 4 m radius at angles of $-40°$, $+40°$ and $0°$ and the microphones consisted of an AKG SA30 stereo pair placed at the centre of the circle and two directional microphones placed 50 cm away from the peripheral loudspeakers. The mixture signals were then computed by adding the reference source image signals together, keeping only the peripheral sources when $J = 2$ or the stereo microphone pair when $I = 2$.

The same algorithms as in section 7.1 were applied, with the difference that the STFT length for FDICA, near-optimal time-invariant filtering and near-optimal local mixing inversion was set to 16384 since this provided the best results for FDICA. Due to the absence of known mixing filters, the complex mixing matrices $\mathbf{A}(f)$ used for near-optimal local mixing inversion were estimated from the reference source images by least squares fitting.

The results are presented in table 2. Oracle estimators span a similar SDR range as in table 1, with the best SDR being achieved by near-optimal time-invariant filtering for the over-determined mixture and by near-optimal local mixing inversion for the under-determined mixture. The SDR difference between near-optimal local mixing inversion and near-optimal binary masking appears smaller than in table 1, which may be due to the fact that the larger window length needed to accurately represent the mixing filters in the frequency domain results in a larger time-frequency overlap of the sources. The comparatively poor performance of FDICA and DUET suggests that better blind algorithms could be found. We emphasize that these results might not be valid for other live microphone recordings, and that more recordings should be made to obtain significant conclusions.

8 Robustness analysis

We have seen in the previous section that the upper performance bounds provided by oracle estimators allow us to evaluate quickly the potential performance of a class of separation algorithms. Nevertheless, blind algorithms may fail to reach these bounds for several reasons. Besides the difficulty of designing an appropriate objective function and optimizing it in a blind context, one of these reasons may be that the separating function is not robust enough, so that performance decreases very quickly as soon as separation parameters are slightly different from the oracle parameters. To measure this robustness, we performed a series of experiments illustrating how oracle estimators can also be used to assess the sensitivity of the separation performance to inaccurate estimation of the oracle parameters.

In addition to the original set of mixing filters described in section 3 for RT = 250 ms, three sets of filters were built similarly by modifying the direction of source 1 to $-38°$, $-36°$ and $-32°$. This amounts to a respective error of $2°$, $4°$ and $8°$ compared to the original direction of $-40°$. Four sets of determined two-source mixtures were obtained by applying these mixing filters to the speech and music sources of section 3. Oracle separation parameters were then learnt on the mixture signals involving modified source directions and applied to the corresponding mixture signals involving the original source direction. One can think of these experiments either as simulating the effect of the movement of a source in the recording room, or as reflecting uncertainty in the estimation of the mixing filters. The results will indicate how the performance degrades when an imperfect estimate is used instead of the oracle one.

8.1 Robustness of multichannel time-invariant filtering

Figure 9 displays the results for the time-domain time-invariant filtering estimator with various hypothesized positions of source 1 for which the oracle filters were learnt. When the true source location is known, the longer the oracle filter, the better the performance. When an erroneous source location is hypothesized, the performance of short oracle filters of less than a hundred taps remains lower than that of longer filters but is reasonably robust. On the contrary, the performance of longer filters decreases by up to 10 dB. When the error on the source direction reaches $8°$, performance is maximized for a filter length of about 500 samples (25 ms) and decreases when longer filters are used. Further investigation show that the SDR values for both estimated sources are similar, but that the distortion is dominated by spatial mislocation for source 1 and by interference for source 2.

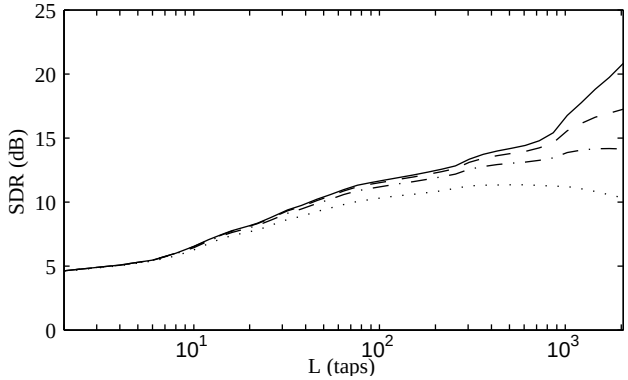

Figure 9. Average performance of the time-invariant filtering oracle on determined two-source mixtures with reverberation time RT = 250 ms as a function of the length L of the demixing filters. Each curve corresponds to a different hypothesized direction of source 1 for which the oracle demixing filters were learnt (plain: true direction, dashed: 2° error, dash-dotted: 4° error, dotted: 8° error).

Previously other authors [13] have performed similar experiments with slightly shorter mixing filters involving only a few nonzero taps and corresponding to a reverberation time RT = 100 ms. In contrast with our approach, they computed (very short) non optimal demixing filters by truncating the adjoint of the mixing system to retain only between 1 and 11 nonzero taps. Their conclusion was that short demixing filters of one or two nonzero taps should be preferred to longer ones, because their increased robustness globally improved SIR when erroneous source locations were hypothesized. Our experiments do not confirm this conclusion: even in the event of a 8° error on the hypothesized source direction, demixing filters of a few hundred to a thousand taps still provide a significant performance improvement over short filters. This indicates that the method used to generate the test demixing filters is important in determining robustness.

8.2 Robustness of local mixing inversion

Figure 10 displays the average performance of the near-optimal local mixing inversion estimator with a free number of active sources per time-frequency point for different hypothesized positions of source 1. Again, performance is more robust, although globally poorer, for short STFT windows up to a thousand samples. For longer windows, errors in the hypothesized source direction decrease the performance by up to 25 dB compared to the oracle performance in ideal conditions. When the error on the source direction reaches 8°, performance is maximized for a window length of about 2400 samples (110 ms). Again, detailed results show that the distortions on the two estimated sources are of different nature but that the SDR values are similar.

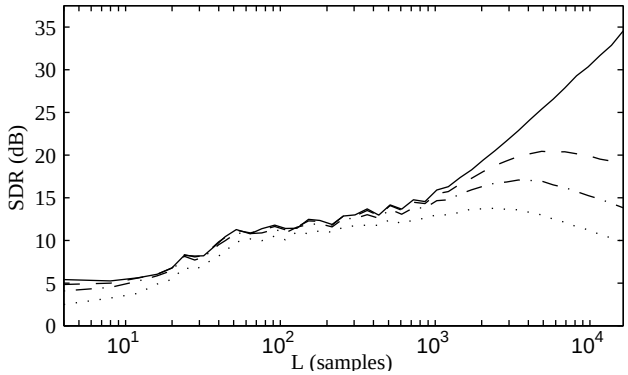

Figure 10. Average performance of near-optimal local mixing inversion with a free number of active sources $J'(n, f)$ for each (n, f) on determined two-source mixtures with reverberation time RT $= 250$ ms as a function of the STFT length L. Each curve corresponds to a different hypothesized direction of source 1 for which the oracle activity patterns were learnt (plain: true direction, dashed: 2° error, dash-dotted: 4° error, dotted: 8° error).

9 Conclusion

In this article, we introduced oracle estimators as a new framework for the benchmarking of source separation algorithms. These estimators can be used to provide upper bounds on the performance of existing and future blind algorithms, choose the best class of algorithms for a given mixture signal or quantify the difficulty of separating this signal. We described explicit oracle estimators for three particular classes of algorithms: multichannel time-invariant filtering, single-channel time-frequency masking and multichannel time-frequency masking.

The study of these oracle estimators on a set of audio mixtures led to three different kinds of conclusions. Firstly, we confirmed and extended the results of previous performance studies based on other types of estimators, such as blind algorithms or near-optimal estimators. In particular, we showed that convolutive mixing with reverberation time RT $= 250$ ms can decrease the maximal performance of time-invariant filtering on determined mixtures by up to 20 dB compared to anechoic mixing (section 4.3), and that the best window length for single-channel time-frequency masking equals about 55 ms for speech and 190 ms for music (section 5.3). Secondly, we reported a few results markedly different from those of previous studies. For instance, we established that the choice of $J' = 2$ active sources per time-frequency point actually increases the maximal performance of local mixing inversion on reverberant mixtures by 7 dB compared to that of a single active source, provided the window length is large enough (section 6.4). We also showed that, even in case of source movements, demixing filters of about a thousand taps remain preferable to shorter filters for the separation of reverberant mixtures (sec-

tion 8.1). Thirdly, we were able to quantify the dependence of oracle demixing filters on the source signals (section 4.2) and the effect of frequency-domain implementation of time-invariant filtering (section 4.4), both of which had not been measured before. We believe that the use of rigorous oracle estimators in our experiments reinforces the validity of these conclusions, when compared to previous studies relying instead on specific blind algorithms or near-optimal estimators.

Comparison of the estimators showed that oracle local mixing inversion with a free number of active sources per time-frequency point outperformed both time-invariant filtering and binary masking on determined and under-determined convolutive mixtures, by at least 4 dB and 8 dB respectively (section 7.3). This suggests that performance advances in blind source separation may be possible by developing blind local mixing inversion algorithms relaxing the common assumption of a single active source per time-frequency point.

The oracle estimators described in this article have been implemented in Matlab and distributed under the GNU Public License as part of the BSS Oracle toolbox at `http://bass-db.gforge.inria.fr/bss_oracle/`.

We are currently considering several directions for future work. Firstly, we plan to derive additional oracle estimators based on the ones considered here. For example, it would be interesting to define a joint near-optimal estimator of the mixing matrix and the source activity patterns for local mixing inversion, to evaluate the importance of having optimal mixing coefficients. Secondly, we hope to increase the relevance of the results for audio mixtures by computing modified oracle estimators using perceptually motivated distortion measures or constraining the amount of musical noise artifacts to lie below a certain acceptability threshold. Finally, we are considering using the results of oracle estimators as training data for the design of new blind separation algorithms. For instance, the observation of oracle source patterns for time-frequency masking could give an insight into which objective functions would be suitable for their blind estimation.

Acknowledgment

The authors are grateful to Scott Rickard and Nikolaos Mitianoudis for providing implementations of DUET and FDICA respectively and to Cédric Févotte and the sound engineers at IRCAM for participating in the recording of real music mixtures. The authors also wish to thank the anonymous reviewers for their careful comments which helped improve the presentation of this article.

References

[1] P. Smaragdis, Blind separation of convolved mixtures in the frequency domain, Neurocomputing 22 (1998) 21–34.

[2] J.-F. Cardoso, Blind source separation : statistical principles, Proceedings of the IEEE 9 (10) (1998) 2009–2025.

[3] D.-T. Pham, J.-F. Cardoso, Blind separation of instantaneous mixtures of non stationary sources, IEEE Trans. on Signal Processing 49 (9) (2001) 1837–1848.

[4] M. J. Reyes-Gomez, B. Raj, 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.

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

[6] L. Benaroya, L. McDonagh, F. Bimbot, 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.

[7] Ö. Yilmaz, S. T. Rickard, Blind separation of speech mixtures via time-frequency masking, IEEE Trans. on Signal Processing 52 (7) (2004) 1830–1847.

[8] N. Roman, D. Wang, G. J. Brown, Speech segregation based on sound localization, Journal of the Acoustical Society of America 114 (4) (2003) 2236–2252.

[9] E. Vincent, Musical source separation using time-frequency source priors, IEEE Trans. on Audio, Speech and Language Processing 14 (1) (2006) 91–98.

[10] S. Araki, R. Mukai, S. Makino, T. Nishikawa, H. Saruwatari, The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech, IEEE Trans. on Speech and Audio Processing 11 (2) (2003) 109–116.

[11] K. E. Hild II, D. Erdogmus, J. C. Principe, Experimental upper bound for the performance of convolutive source separation methods, IEEE Trans. on Signal Processing 54 (2) (2006) 627–635.

[12] A. Westner, V. M. Bove, Blind separation of real world audio signals using overdetermined mixtures, in: Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA), 1999, pp. 251–256.

[13] R. V. Balan, J. P. Rosca, 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.

[14] M. Hofbauer, On the FIR inversion of an acoustical convolutive mixing system: properties and limitations, in: Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA), 2004, pp. 643–651.

[15] K. E. Hild II, R. Jensen, Experimental upper bound for convolutive mixtures of speech by maximizing SIR, in: Proc. IEEE Int. Workshop on Machine Learning for Signal Processing (MLSP), 2006, pp. 235–240.

[16] K. Matsuoka, S. Nakashima, Minimal distortion principle for blind source separation, in: Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA), 2001, pp. 722–727.

[17] N. Mitianoudis, M. E. Davies, Audio source separation of convolutive mixtures, IEEE Trans. on Speech and Audio Processing 11 (5) (2003) 489–497.

[18] T. Takatani, T. Nishikawa, H. Saruwatari, K. Shikano, SIMO-model-based independent component analysis for high-fidelity blind separation of acoustic signals, in: Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA), 2003, pp. 993–998.

[19] D. Schobben, K. Torkkola, P. Smaragdis, Evaluation of blind signal separation methods, in: Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA), 1999, pp. 261–266.

[20] E. Vincent, R. Gribonval, C. Févotte, Performance measurement in blind audio source separation, IEEE Trans. on Audio, Speech and Language Processing 14 (4) (2006) 1462–1469.

[21] S. van de Par, A. Kohlrausch, G. Charestan, R. Heusdens, A new psycho-acoustical masking model for audio coding applications, in: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2002, pp. II–1805–1808.

[22] U. P. Svensson, U. R. Kristiansen, Computational modelling and simulation of acoustic spaces, in: Proc. AES 22nd Conf. on Virtual, Synthetic and Entertainment Audio, 2002, pp. 1–20.

[23] J. Nocedal, S. J. Wright, Numerical optimization, Springer, New York, NY, 1999.

[24] D. W. Griffin, J. S. Lim, Signal estimation from modified short-time Fourier transform, IEEE Trans. on Acoustics, Speech and Signal Processing 32 (2) (1984) 236–243.

[25] C. Servière, Separation of speech signals with segmentation of the impulse responses under reverberant conditions, in: Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA), 2003, pp. 511–516.

[26] J. P. Princen, A. B. Bradley, Analysis/synthesis filter bank design based on time domain aliasing cancellation, IEEE Trans. on Acoustics, Speech and Signal Processing 34 (5) (1986) 1153–1161.

[27] S. Araki, S. Makino, H. Sawada, R. Mukai, Reducing musical noise by a fine-shift overlap-add method applied to source separation using a time-frequency mask, in: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2005, pp. III–81–84.

[28] R. Gribonval, Piecewise linear source separation, in: Proc. SPIE, Vol. 5207 Wavelets: Applications in Signal and Image Processing X, 2003, pp. 297–310.

[29] D. Kolossa, 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.

[30] A. Aïssa-El-Bey, K. Abed-Meraim, Y. Grenier, Underdetermined blind source separation of audio sources in time-frequency domain, in: Proc. Workshop on Signal Processing with Sparse/Structured Representations (SPARS), Rennes, France, 2005, pp. 67–70.

[31] P. Bofill, M. Zibulevsky, Underdetermined blind source separation using sparse representations, Signal Processing 81 (2001) 2353–2362.

[32] J. P. Rosca, C. Borss, 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] S. Winter, H. Sawada, S. Makino, On real and complex-valued l_1-norm minimization for overcomplete blind source separation, in: Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2005, pp. 86–89.

[34] R. H. Lambert, Difficulty measures and figures of merit for source separation, in: Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA), 1999, pp. 133–138.