

The logo for the Centre for Digital Music features a horizontal bar with a green-to-blue gradient and vertical lines. The text "centre for digital music" is written in white, lowercase letters across the bar.

centre for digital music

Oracle Estimators for the Benchmarking of Source Separation Algorithms

EMMANUEL VINCENT, RÉMI GRIBONVAL AND MARK D. PLUMBLEY

Technical Report C4DM-TR-06-03
28 July 2006

Oracle Estimators for the Benchmarking of Source Separation Algorithms

Emmanuel Vincent¹, Rémi Gribonval² and Mark D. Plumbley¹

¹Centre for Digital Music
Queen Mary, University of London
Mile End Road – London E1 4NS – United Kingdom
emmanuel.vincent@elec.qmul.ac.uk

²METISS group
IRISA-INRIA
Campus de Beaulieu – 35042 Rennes CEDEX – France
remi.gribonval@irisa.fr

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 four particular classes of algorithms: beamforming, single-channel time-frequency masking, multichannel time-frequency masking and best basis masking. We evaluate their performance on various audio mixtures and study their robustness. We draw several conclusions regarding the performance bounds of blind algorithms, the choice of the best class of algorithms and the assessment of 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.

Keywords: blind source separation, performance, evaluation, benchmark.

1 Introduction

Most audio, video and biomedical signals are mixtures of several sources that are active simultaneously. In general, the mixing process consists of a nonlinear time-varying transform applied to the source signals. However, in most cases, it can be modeled as a time-invariant linear filtering process. The i th channel of the observed mixture ($1 \leq i \leq I$) is then expressed as

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

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 involve additional (possibly fractional) delays, 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 time-invariant *beamforming*, 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:

- 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 beamforming 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 the performance of these estimators leads to three main applications: providing theoretical upper bounds on the performance of existing blind algorithms, predicting the adequacy of different classes of algorithms for a given mixture signal and quantifying the difficulty of separating it.

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 works in [11, 12, 13, 14], which focused on determining near-optimal demixing filters for determined or over-determined convolutive mixtures by inverting the mixing filter system or minimizing the energy of interference. Strictly speaking, the filters developed by these authors are not oracle estimates, since the criterion which they optimize is not equal to the chosen performance measure [14]. In the following, we provide a rigorous framework for the definition of oracle estimators and we compare the resulting oracle demixing filters to some of these near-optimal demixing filters. We also design oracle estimators for three other classes of algorithms, namely single-channel time-frequency masking, multichannel time-frequency masking and best basis 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 simulated audio recordings and synthetic mixtures used throughout the article. Then, in sections 4, 5, 6 and 7, we present oracle estimators for four different classes of source separation algorithms and evaluate their performance under various constraints. We detail the various applications of these estimators in section 8, including their comparison with some usual blind separation algorithms, and we analyze their robustness in section 9. Finally, we summarize the contributions of this article in section 10 and point out further research directions.

2 Oracle *vs.* blind source separation

2.1 Principle

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

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

where f is a fixed function, θ a vector of parameters depending on 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 time-invariant beamforming, 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 $d(\hat{\mathbf{y}}, \mathbf{y})$. We define the *oracle estimator* $\tilde{\theta}(\mathbf{y}, \mathbf{x}, \Theta)$ to be the vector of parameters resulting in the smallest distortion among the set of acceptable parameters Θ :

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

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

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

2.2 Examples

In practice, the target may assume different forms. For example, it may be a single-channel signal, such as the j th source signal $s_j(t)$ or the *image* of the j th source on the i th mixture channel, defined by [15]

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

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 $\mathbf{s}(t) = [s_1(t), \dots, s_J(t)]^T$ or the images of all the sources on all channels $\mathbf{s}^{\text{img}}(t) = [s_{1,1}^{\text{img}}(t), \dots, s_{1,J}^{\text{img}}(t), \dots, s_{I,1}^{\text{img}}(t), \dots, s_{I,J}^{\text{img}}(t)]^T$. The latter is a common target for blind source separation algorithms [16, 17]. Indeed, these algorithms can estimate the source signals only up to an arbitrary scaling [2] or filtering [1], but these indeterminacies disappear when considering the source images instead [15, 16]. Moreover, reference source images for realistic mixtures can easily be acquired using simulated recording experiments [18, 14]. 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, except in section 4.3.3 where we conduct a comparison with previous studies using $\mathbf{y}(t) = \mathbf{s}(t)$.

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

In the following, we measure the distortion $d(\hat{\mathbf{y}}, \mathbf{y})$ between a target and its estimate using the Euclidean distortion measure

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

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 Source-to-Distortion Ratio (SDR) expressed in decibels (dB)

$$\text{SDR} = 10 \log_{10} \frac{\|\mathbf{y}\|^2}{\|\hat{\mathbf{y}} - \mathbf{y}\|^2}. \quad (7)$$

Minimizing the distortion $d(\hat{\mathbf{y}}, \mathbf{y})$ is equivalent to maximizing the SDR. The definition of the SDR incorporates all possible kinds of distortion arising from different source separation algorithms, including interference from other sources, nonlinear artifacts, filtering distortion and spatial distortion of the target signal [19]. The Source-to-Interference Ratio (SIR) criterion used in some studies on beamforming algorithms [14] is less relevant in a general context, since it does not measure distortions other than interference and can provide a high performance rating despite strong distortion of the target. Moreover, contrary to SDR, SIR does not lead to exact closed form expressions for simple oracle estimators, so that only near-optimal estimators can be computed instead [14]. 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 [20].

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

We collected ten multitrack music recordings (*master tapes*) from different genres¹, each containing three sources playing synchronously and in harmony. We also selected speech sources from thirty English speakers from as many different audio books². 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 [21] using the Roomsim toolbox³. 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 average delay of 64 samples between the loudspeakers and the microphones was compensated, so that the absolute delay between the sources and the first two mixture channels was smaller than 5 samples. The length of the mixing filters was assessed using the reverberation time RT defined to be the average lag for which the magnitude of the mixing filters becomes less than 1/1000 of its maximal magnitude. This quantity is typically on the order of 250 ms in a small quiet meeting room and 2 s in a large 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).

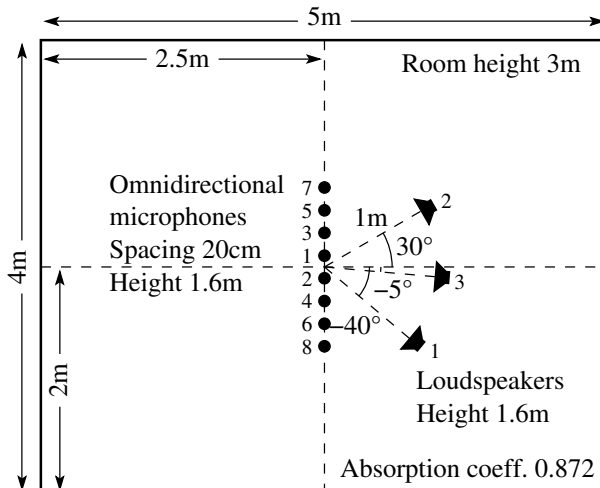


Figure 1: Microphone and loudspeaker positions for simulated room recordings with $RT = 50$ ms. Anechoic recordings were generated using the same configuration with an absorption coefficient of 1. Recordings with $RT = 250$ ms and $RT = 1.25$ s were simulated using the same microphone and loudspeaker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

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

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

In the experiments of sections 4 to 7 we measure the average performance trends of the oracle estimators applied to this data. The reported values of SDR could be slightly different on other data. We provide a more detailed account for this using confidence bounds in the experiments of section 8.

¹These 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.

²These were chosen among public domain audio books from <http://librivox.org/>

³<http://media.paisley.ac.uk/~campbell/Roomsim/>

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

4 Beamforming oracle

Source separation algorithms for determined or over-determined mixtures, *e.g.* [2, 1, 16, 17, 4], are generally based on beamforming, which relies mainly on the spatial diversity of the sources and to a lesser degree on their spectral diversity. In the following, we restrict ourselves to time-invariant beamforming, which is the process of filtering the mixture channels by 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 [14]. When these filters are carefully chosen, beamforming attenuates sounds coming from certain spatial directions and at certain frequencies where interference dominates.

4.1 Definition of the separating function

Beamforming 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 4.3.4. 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

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

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

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

Some algorithms derive the image demixing filters $w_{ijk}(\tau)$ from the source demixing filters $w_{jk}(\tau)$ by matrix computation [16]. Other algorithms estimate unconstrained image demixing filters $w_{ijk}(\tau)$ directly using a cost function to assess the quality of the reconstructed mixture channels [17].

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

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

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 $\tilde{w}_{c\eta}$ 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 , $1 \leq \eta \leq D$ [22]. 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 $\tilde{\mathbf{w}}_c = [\tilde{w}_{c,1}, \dots, \tilde{w}_{c,D}]^T$ is equal to

$$\tilde{\mathbf{w}}_c = \mathbf{G}^{-1} \mathbf{r}_c \quad (12)$$

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 $r_{c\eta} = \langle y_c, x_\eta \rangle$.

4.3 Examples

The performance of separation algorithms based on beamforming depends on various factors, including the number of mixture channels, the length of the mixing filters and the constraints over the demixing filters. The oracle estimator above can provide some insights about these issues.

4.3.1 Effect of reverberation time

In theory, determined or over-determined mixtures can be perfectly separated using beamforming by constructing the demixing filter system to be the pseudo-inverse of the mixing filter system. In practice however, the optimal demixing filters are often much too long and shorter suboptimal filters are used instead. We evaluated the performance of oracle demixing filters on determined two-source mixtures by varying the reverberation time RT . Figure 2 shows that beamforming 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 on realistic reverberant mixtures, 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 beamforming can reject interference from at most one direction at low frequencies in a two-channel mixture [10]. Thus virtual interference sources generated by reverberation at random spatial directions cannot be perfectly cancelled at low frequencies whatever filter length is used.

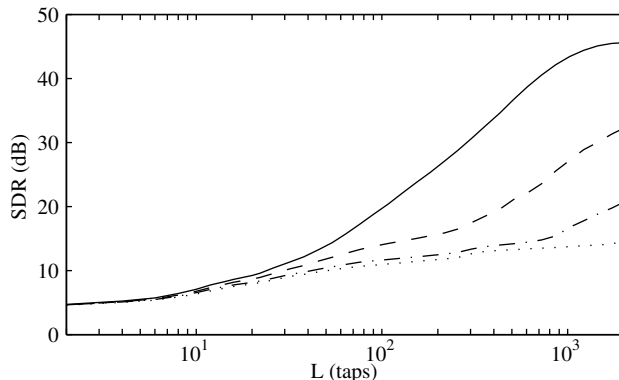


Figure 2: Average performance of the beamforming oracle on determined two-source mixtures as a function of the length L of the demixing filters. Each curve corresponds to a different reverberation time (plain: anechoic, dashed: $RT = 50$ ms, dash-dotted: $RT = 250$ ms, dotted: $RT = 1.25$ s).

4.3.2 Effect of the number of channels

When the number of mixture channels is increased, demixing filters can have more complex spatial responses and reject interference from several unrelated directions 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 [11, 12]. Note that, in the present case, the SDR obtained with oracle filters of length L on I mixture channels is always larger (up to 7 dB) than the SDR obtained with oracle filters of length $LI/2$ on two channels, despite the fact that both demixing systems have the same total number of coefficients.

4.3.3 Filter dependence on the source spectra

Previous studies on the performance of beamforming algorithms derived near-optimal source demixing filters by computing the pseudo-inverse of the mixing filter system [11, 13]. These filters can easily be shown to be identical to the oracle demixing filters for uncorrelated white noise sources. However oracle filters provide a better performance in general by taking into account the power spectral densities of the sources. Hence

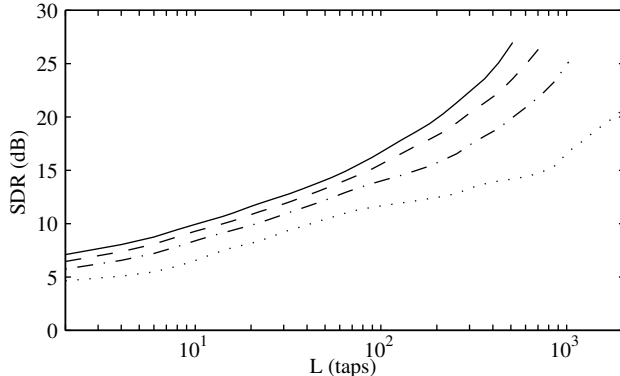


Figure 3: Average performance of the beamforming oracle on over-determined two-source mixtures with reverberation time $RT = 250$ ms as a function of the length L of the demixing filters. Each curve corresponds to a different number of mixture channels (plain: $I = 8$, dashed: $I = 6$, dash-dotted: $I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

they are able to cancel a larger proportion of the interference sources at the frequencies where they have more energy and allow a larger relative distortion of the target source at the frequencies where it has little energy. To quantify the importance of this issue, we compared oracle source demixing filters with pseudo-inverse filters determined as in [13] on determined two-source mixtures. The results are plotted in figure 4. Pseudo-inverse filters provide a similar performance for music mixtures and speech mixtures. However oracle filters perform consistently better than pseudo-inverse filters, with an average SDR improvement of 2 dB for speech and 4 dB for music. For long demixing filters, the SDR obtained with oracle filters becomes 5 dB larger for music than for speech for the considered data.

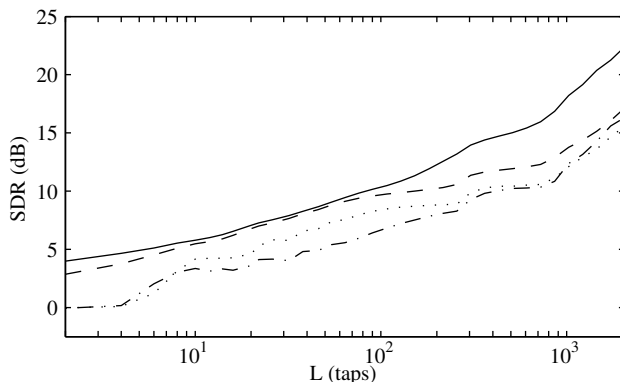


Figure 4: Average performance of various source demixing filters as a function of their length L on determined two-source mixtures with reverberation time $RT = 250$ ms. Plain: oracle filters applied to music data. Dashed: oracle filters applied to speech data. Dash-dotted: pseudo-inverse filters applied to music data. Dotted: pseudo-inverse filters applied to speech data.

4.3.4 Frequency-domain approximate implementation

For computational convenience, source separation algorithms often implement beamforming 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 $S_{ij}^{\text{img}}(n, f)$ of the sources and the source images respectively are then estimated by $\hat{S}_j(n, f) = \sum_{k=1}^I w_{jk}(f)X_k(n, f)$ and $\hat{S}_{ij}^{\text{img}}(n, f) = \sum_{k=1}^I w_{ijk}(f)X_k(n, f)$ 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 [23]. As explained in [24], this formulation is not exactly equivalent to time-domain beamforming, 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 compute 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 a tighter upper bound on the performance of many blind algorithms.

The SDR curves for source images, plotted in figure 5, show that the time-domain oracle performs 2 dB better than its frequency-domain counterpart on average, 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 beamforming in the frequency domain, such as the one proposed in [24].

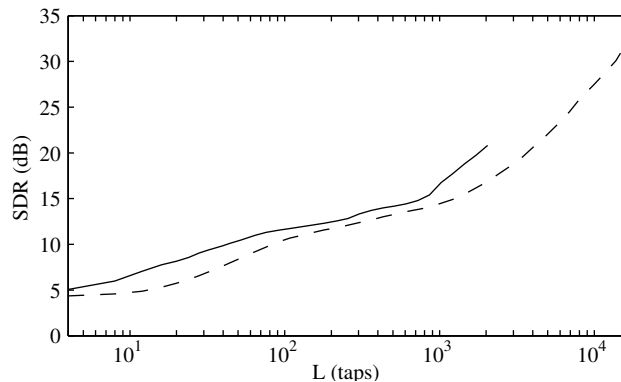


Figure 5: Average performance of various source image demixing filters on determined two-source mixtures with reverberation time $RT = 250$ ms. Plain: time-domain oracle demixing filters of length L . Dashed: near-optimal frequency-domain demixing filters with L frequency bins. Long oracle filters could not be estimated due to large memory requirements.

5 Single-channel time-frequency masking oracle

Beamforming 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 5.3.3 and assume instead that time-frequency masking is performed using an orthonormal time-frequency basis, such as the Modified Discrete Cosine Transform (MDCT) [25]. 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 $\hat{s}_j^{\text{img}}(t)$ can be written as

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

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]. We also distinguish *unconstrained masks* from *constrained masks* whose coefficients verify the unitary sum constraint

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

This constraint is often used in standard mask estimation algorithms such as adaptive Wiener filtering [6]. Relaxing it potentially improves performance by allowing a larger range of masks. However blind estimation of unconstrained masks is more difficult in practice since it requires more information about the correlations between sources in each time-frequency point.

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 $\hat{y}_c(t) = \sum_{m=1}^M \epsilon_{cm} \langle x, \phi_m \rangle \phi_m(t)$. Since MDCT elements are orthonormal, the total Euclidean distortion over the target can be decomposed as $\|\hat{\mathbf{y}} - \mathbf{y}\|^2 = \sum_{m=1}^M (\epsilon_{cm} \langle x, \phi_m \rangle - \langle y_c, \phi_m \rangle)^2$, yielding after simple computation

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

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.

When no constraint is set over the sum of the masks, the oracle binary masking coefficients are found by

$$\tilde{\epsilon}_{cm} = \begin{cases} 0 & \text{if } r_{cm} < \frac{1}{2}, \\ 1 & \text{otherwise,} \end{cases} \quad (16)$$

and the oracle real-valued masking coefficients by

$$\tilde{\epsilon}_{cm} = \begin{cases} 0 & \text{if } r_{cm} < 0, \\ 1 & \text{if } r_{cm} > 1, \\ r_{cm} & \text{otherwise.} \end{cases} \quad (17)$$

When the masks are subject to the unitary sum constraint (14), the oracle binary masking coefficients are given by $\tilde{\epsilon}_{c_m m} = 1$ for $c_m = \arg \max_c r_{cm}$ and $\tilde{\epsilon}_{cm} = 0$ for $c \neq c_m$. Under this constraint, the computation of the oracle real-valued masking coefficients becomes 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 [22]. When the number of sources is large, it can also be obtained more efficiently using iterative active set or interior point methods [22].

5.3 Examples

The performance of time-frequency masking algorithms depends on various parameters, in particular on the MDCT length and on the use of binary or real-valued masks. Oracle estimators provide a natural framework to help us choose relevant parameters depending on the context. The following experiments were done using MDCT bases built from sine windows [25].

5.3.1 Choice of the length of MDCT elements

Time-frequency masking is a nonlinear operation and does not guarantee the preservation of the perceptually important temporal and spectral continuity properties of the sources. Consequently, when the sources overlap in the time-frequency plane, the distortion on the source estimates is often dominated by musical noise artifacts [7]. This distortion can be reduced by choosing the length of the MDCT basis elements so that the sources overlap as little as possible. In figure 6, a comparison of the various constrained masking oracles on single-channel two-source mixtures 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 our experiments, 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.

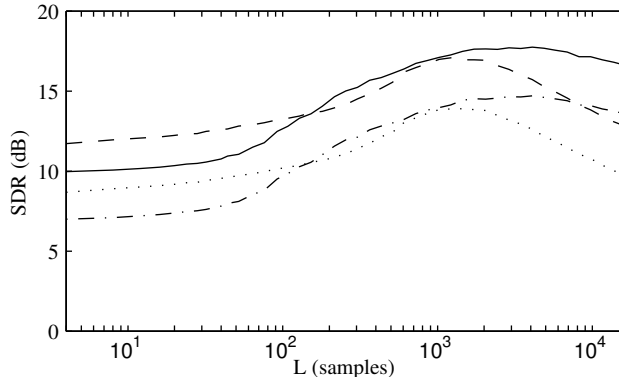


Figure 6: Average performance of the constrained time-frequency masking oracles on single-channel two-source mixtures as a function of the length L of the MDCT basis elements. Plain: real-valued masks applied to music data. Dashed: real-valued masks applied to speech data. Dash-dotted: binary masks applied to music data. Dotted: binary masks applied to speech data.

5.3.2 Relevance of the unitary sum constraint

When the number of sources is equal to two, unconstrained and constrained oracle estimators are equivalent. Indeed, it is easy to show that, given the expression of unconstrained estimators in (16) and (17), the equality $r_{1,m} + r_{2,m} = 1$ stemming from the fact that $s_1^{\text{img}}(t) + s_2^{\text{img}}(t) = x(t)$ implies that unconstrained oracle estimates verify $\tilde{\epsilon}_{1,m} + \tilde{\epsilon}_{2,m} = 1$. However, this is no longer true when the number of sources is three or more. For example, when $J = 3$, $r_{1,m} = r_{2,m} = 1$ and $r_{3,m} = -1$, the unconstrained oracle masking coefficients $\tilde{\epsilon}_{1,m} = \tilde{\epsilon}_{2,m} = 1$ and $\tilde{\epsilon}_{3,m} = 0$ do not verify the unitary sum constraint (14). Figure 7 shows that the loss of performance due to this constraint on single-channel three-source mixtures is smaller than 0.1 dB around the optimal MDCT length. We conclude that this loss is well compensated by the increased ease of estimation of constrained masks in a blind context. This also means that the performance of the constrained real-valued oracle estimator can be approximated much more quickly using the unconstrained oracle estimator.

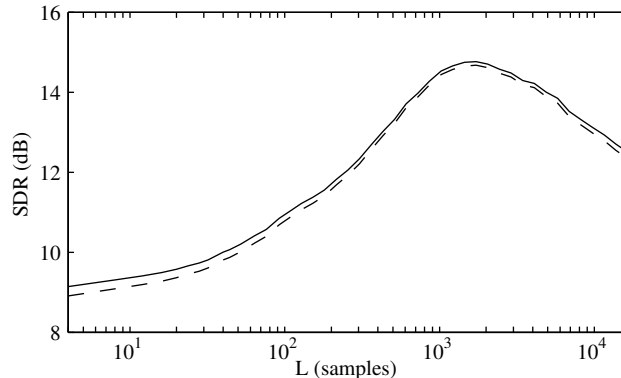


Figure 7: Average performance of the real-valued time-frequency masking oracles on single-channel three-source mixtures as a function of the length L of the MDCT basis elements. Plain: unconstrained masks. Dashed: masks verifying a unitary sum constraint.

5.3.3 Use of 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 [23].

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 obtain 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 $\|\hat{\mathbf{Y}}(n, f) - \mathbf{Y}(n, f)\|^2 = |X(n, f)|^2 \sum_{c=1}^C (\epsilon_c(n, f) - R_c(n, f))^2$ 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 8 shows that the average SDR obtained with near-optimal binary masks 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 [26] 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.

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 beamforming and time-frequency masking by jointly exploiting spatial diversity and time-frequency diversity.

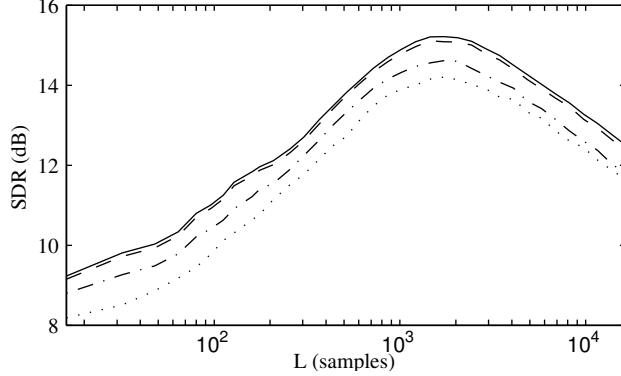


Figure 8: Average performance of the constrained binary MDCT time-frequency masking oracle and the near-optimal STFT time-frequency masks on single-channel two-source mixtures as a function of the length L of the MDCT/STFT elements. Each curve corresponds to a different over-completeness factor (plain: eight-times overcomplete STFT, dashed: four-times overcomplete STFT, dash-dotted: twice overcomplete STFT, dotted: MDCT).

6.1 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 6.3.2. In the following, we focus on two different separating functions: binary masking and local mixing inversion [27, 7]. For the sake of conciseness, alternative separating functions, such as those stemming from the chaining of ICA and binary masking [28], 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

$$\hat{s}_{ij}^{\text{img}}(t) = \sum_{m=1}^M \epsilon_{jm} \langle x_i, \phi_m \rangle \phi_m(t). \quad (18)$$

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 $[\langle x_1, \phi_m \rangle, \dots, \langle x_I, \phi_m \rangle]^T = \mathbf{A}[\langle s_1, \phi_m \rangle, \dots, \langle s_J, \phi_m \rangle]^T$. 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 [27, 29]

$$\begin{cases} \widehat{\langle s_j, \phi_m \rangle} & = 0 & j \notin \mathcal{J}_m, \\ \left[\widehat{\langle s_j, \phi_m \rangle} \right]_{j \in \mathcal{J}_m}^T & = \mathbf{A}_{\mathcal{J}_m}^\dagger [\langle x_i, \phi_m \rangle]_{1 \leq i \leq I}^T \end{cases} \quad (19)$$

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

$$\widehat{\langle s_{j_m}, \phi_m \rangle} = \frac{\mathbf{A}_{j_m}^T [\langle x_i, \phi_m \rangle]_{1 \leq i \leq I}^T}{\|\mathbf{A}_{j_m}\|_2^2}. \quad (20)$$

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

$$\widehat{s}_j(t) = \sum_{m=1}^M \widehat{\langle s_j, \phi_m \rangle} \phi_m(t), \quad (21)$$

$$\widehat{s}_{ij}^{\text{img}}(t) = \sum_{m=1}^M a_{ij} \widehat{\langle s_j, \phi_m \rangle} \phi_m(t). \quad (22)$$

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 [30, 29] 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

$$\tilde{\mathcal{J}}_m = \arg \min_{\mathcal{J}_m \in \mathcal{P}} \sum_{j=1}^J \sum_{i=1}^I \left(\langle \widehat{s}_{ij}^{\text{img}}, \phi_m \rangle - \langle s_{ij}^{\text{img}}, \phi_m \rangle \right)^2 \quad (23)$$

where the role of the activity pattern \mathcal{J}_m in the right hand side is implicit (see (19), (21), (22)). 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 Examples

The performance of multichannel time-frequency masking depends on various parameters, and in particular on the size of the MDCT basis elements and the assumed number of active sources. Oracle estimators provide some insight into the role of these factors on the achievable performance.

6.3.1 Choice of the number of active sources

The assumption that there is a single active source in each time-frequency point allows for simple algorithm design in a blind setting. However, algorithms based on this assumption are known to introduce musical

noise artifacts due to the many zeroes created in the time-frequency estimates of the sources [7]. Allowing a larger number of active sources can potentially reduce these artifacts. 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 9 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 figures 6 and 7 when the length is increased or decreased.

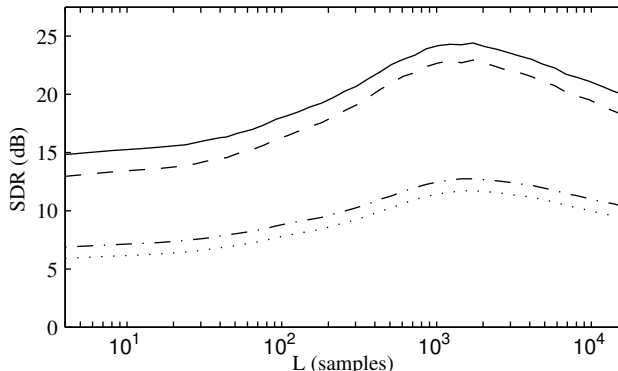


Figure 9: Average performance of the multichannel time-frequency masking oracles on two-channel three-source instantaneous mixtures as a function of MDCT length L . Plain: local mixing inversion with a free number of active sources J'_m for each time-frequency point m . Dashed: $J' = 2$ active sources. Dash-dotted: $J' = 1$ active source. Dotted: binary masking.

6.3.2 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 [31, 32] 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

$$\begin{cases} \widehat{S}_j(n, f) & = 0 & j \notin \mathcal{J}(n, f), \\ \left[\widehat{S}_j(n, f) \right]_{j \in \mathcal{J}(n, f)} & = \mathbf{A}_{\mathcal{J}(n, f)}^\dagger(f) \mathbf{X}(n, f) \end{cases} \quad (24)$$

and the source image STFTs are derived as $\widehat{S}_{ij}^{\text{img}}(n, f) = a_{ij}(f)\widehat{S}_j(n, f)$. Source waveforms are recovered by STFT inversion using the overlap-add method [23].

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 obtain near-optimal binary masks and activity patterns instead by separate minimization of $\|\widehat{\mathbf{S}}^{\text{img}}(n, f) - \mathbf{S}^{\text{img}}(n, f)\|^2$ in each time-frequency point.

Figure 10 illustrates the corresponding performance 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 probably related to the fact that the approximate modeling of the mixing system by a set of complex mixing matrices is more accurate with a large number of frequency bins, *i.e.* for large windows. For $J' = 2$, pseudo-inversion (24) amplifies the effect of the inaccurate modeling for small windows, while for $J' = 1$ this effect is more limited. This may explain the performance degradation observed with $J' = 2$ in a previous study [31], which used short windows of 16 ms.

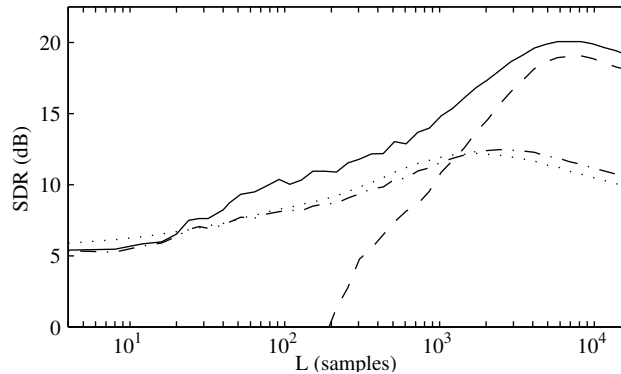


Figure 10: Average performance of multichannel near-optimal time-frequency masks on two-channel three-source mixtures with reverberation time $RT = 250$ ms as a function of STFT length L . Plain: local mixing inversion with a free number of active sources $J'(n, f)$ for each time-frequency point (n, f) . Dashed: $J' = 2$ active sources. Dash-dotted: $J' = 1$ active source. Dotted: binary masking.

7 Best basis masking oracle

We have seen in the previous sections that the choice of the window length of an MDCT basis or a STFT has a substantial role in the resulting performance of oracle separation with single- or multichannel masking. More generally, masking can be performed in any orthonormal basis $\{\phi_m(t)\}_m$, and carefully choosing a basis among a large *library* \mathcal{L} of bases could improve the separation performance. This idea was suggested in [27] and applied recently in [33].

7.1 Definition of the separating function

In each basis $\mathcal{B}^\lambda = \{\phi_m^\lambda(t)\}_m$ of a large library $\mathcal{L} = \{\mathcal{B}^\lambda\}_\lambda$ of orthonormal bases, separation is performed just as expressed in (13) for single-channel binary or real-valued masking, in (18) for multichannel binary masking and in (19), (21) and (22) for multichannel local mixing inversion. The main difference with the previous sections is that the basis index λ is now a parameter of the separating function, in addition to the masking coefficients ϵ_{jm} or the activity patterns \mathcal{J}_m .

7.2 Computation of the oracle parameters

For a fixed basis \mathcal{B}^λ , oracle masking parameters $\tilde{\epsilon}_{jm}^\lambda$ or $\tilde{\mathcal{J}}_m^\lambda$ can be chosen for each basis element separately according to (16), (17) or (22) so as to minimize the distortion among all admissible masking parameters. One can easily check that each oracle masking parameter only depends on the target $\mathbf{y}(t)$, the mixture $\mathbf{x}(t)$

and the corresponding basis element $\phi_m^\lambda(t)$, and that the total oracle distortion achieved with the basis \mathcal{B}^λ is expressed as

$$\tilde{d}(\mathbf{y}, \mathbf{x}, \mathcal{B}^\lambda) = \sum_{m=1}^M \tilde{d}(\mathbf{y}, \mathbf{x}, \phi_m^\lambda). \quad (25)$$

Therefore, the computation of the oracle basis and the associated masking parameters can be performed in two steps: first computation of the oracle masking parameters for each element m of each basis λ while keeping track of the corresponding distortion $\tilde{d}(\mathbf{y}, \mathbf{x}, \phi_m^\lambda)$, then selection of the best basis index $\tilde{\lambda} = \arg \min_{\lambda} \tilde{d}(\mathbf{y}, \mathbf{x}, \mathcal{B}^\lambda)$. This approach is combinatorial in the general case. However, since the Euclidean distortion measure satisfies the additivity property given in [34], efficient algorithms exist when the library has a hierarchical structure such that different bases share common elements [34]. The most classical examples of such libraries are the local Cosine Packet (CP) and Wavelet Packet (WP) libraries [35].

7.3 Examples

The performance of best basis masking depends on various parameters, in particular on the structure of the library of bases. CP and WP bases are defined by their minimum and maximum packet depths D_{\min} and D_{\max} , in addition to the bell type for CP and the wavelet filter type for WP. For CP bases, D_{\min} and D_{\max} determine the minimum and maximum length of analysis windows, which are equal to $T \times 2^{-D_{\max}}$ and $T \times 2^{-D_{\min}}$ samples, with adjacent windows overlapping on $T \times 2^{-D_{\max}-1}$ samples, where T is the length of the signals. For WP bases, D_{\min} and D_{\max} determine the minimum and maximum bandwidth of frequency subbands into which the analyzed signal is filtered, which are equal to $F_s \times 2^{-D_{\max}}$ and $F_s \times 2^{-D_{\min}}$, where F_s is the sampling frequency. Oracle estimators can be used to better understand the role of these factors. In the following, the Wavelab toolbox⁴ was used to compute CP and WP transforms and find the oracle basis given distortion values. CP bases were built with a sine bell, and WP bases with a ‘symmlet-8’ filter.

7.3.1 Choice of the minimal and maximal packet depth

Comparing experiments on single-channel two-source mixtures presented in figures 11 and 12 with previous experiments in figures 6 and 7, it appears that oracle bases do not improve performance as much as expected compared to fixed-size MDCT bases, even when optimal minimum and maximum tree depths are considered. More precisely, the average SDR using MDCT with an optimal length (1800 samples or 80 ms) being equal to 17.3 dB for real-valued masking and 14.2 dB for binary masking, oracle CP bases provide respective SDR improvements of 1.1 dB and 0.9 dB for these two types of masks, while oracle WP bases result in SDR deteriorations of 0.1 dB and 0.4 dB. The optimal value of D_{\max} equals 9 for CP bases, corresponding to a minimal window length of 1024 samples (45 ms), and 18 for WP bases, which means that some oracle WP bases contain some elements with the narrowest possible bandwidth. In theory, the optimal value of D_{\min} equals 0 both for CP and WP bases, since it leads to the largest possible library of bases. However near-optimal performance can be achieved more quickly using larger values of D_{\min} . For example, the best performance of CP bases with $D_{\max} = 9$ decreases by 0.04 dB only with $D_{\min} = 6$ instead of $D_{\min} = 0$, corresponding to a maximal window length of 4608 samples (210 ms) instead of 2^{18} samples (11.9 s).

7.3.2 Computation of the oracle generic WP basis

Given a family of mixture and target signals $\mathbf{x}_l(t)$ and $\mathbf{y}_l(t)$, it is possible to efficiently select a single basis providing the best performance on average. This oracle generic basis is obtained by solving the optimization problem

$$\tilde{\lambda} = \arg \min_{\lambda} \sum_l \tilde{d}(\mathbf{y}_l, \mathbf{x}_l, \mathcal{B}^\lambda) = \arg \min_{\lambda} \sum_{m=1}^M \left(\sum_l \tilde{d}(\mathbf{y}_l, \mathbf{x}_l, \phi_m^\lambda) \right). \quad (26)$$

A WP basis essentially corresponds to a non-uniform recursive partitioning of the frequency axis, and can be described by a binary tree [35]. Figures 13 and 14 show the tree structure of the oracle generic WP basis

⁴<http://www-stat.stanford.edu/~wavelab/>

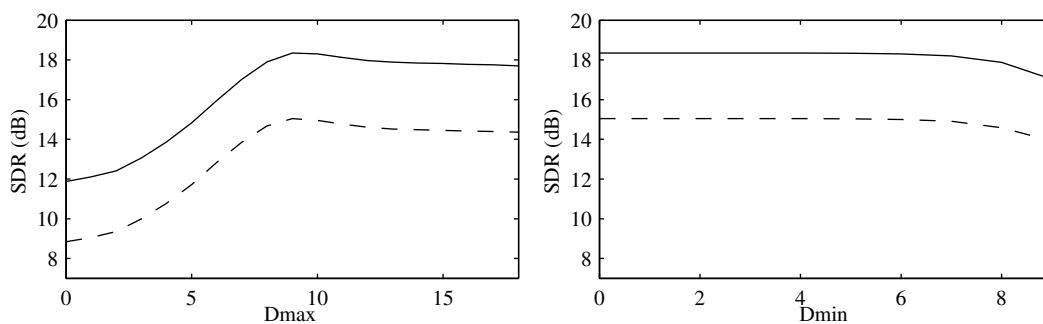


Figure 11: Average performance of the constrained best CP basis masking oracle on single-channel two-source mixtures. Left: SDR as a function of the maximal packet depth D_{\max} with minimal depth $D_{\min} = 0$. Right: SDR as a function of the minimal packet depth D_{\min} with maximal depth $D_{\max} = 9$. Plain: real-valued masking. Dashed: binary masking.

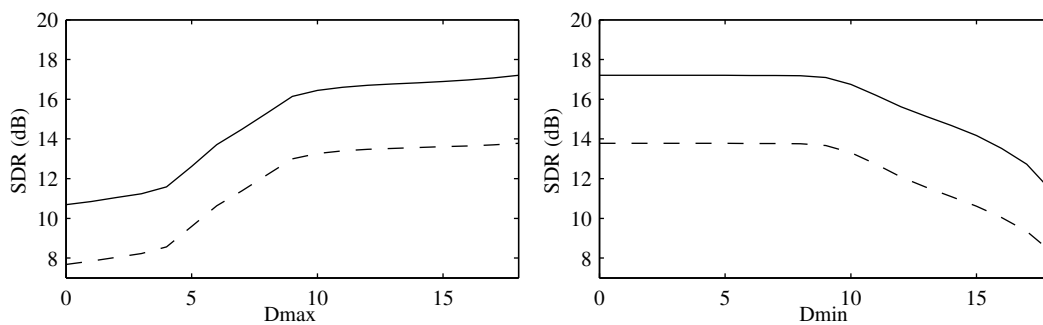


Figure 12: Average performance of the constrained best WP basis masking oracle on single-channel two-source mixtures. Left: SDR as a function of the maximal packet depth D_{\max} with minimal depth $D_{\min} = 0$. Right: SDR as a function of the minimal packet depth D_{\min} with maximal depth $D_{\max} = 18$. Plain: real-valued masking. Dashed: binary masking.

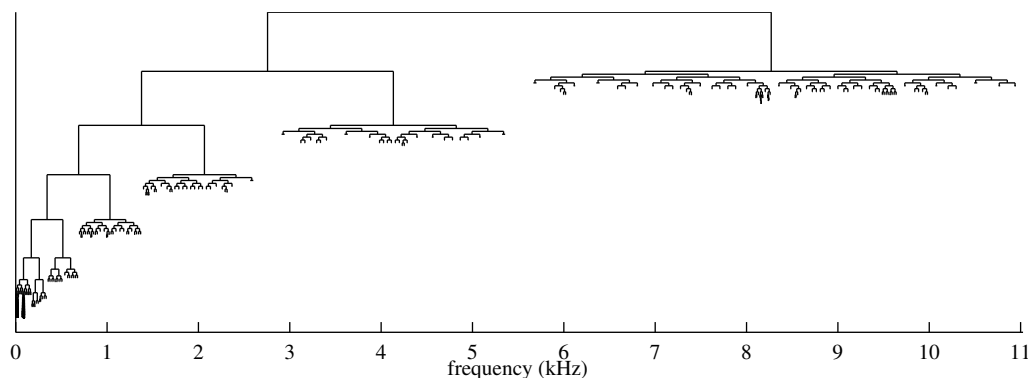


Figure 13: Tree representation of the oracle generic WP basis for real-valued masking of speech mixtures. The height of each branch is proportional to the decrease in distortion \tilde{d} obtained by splitting.

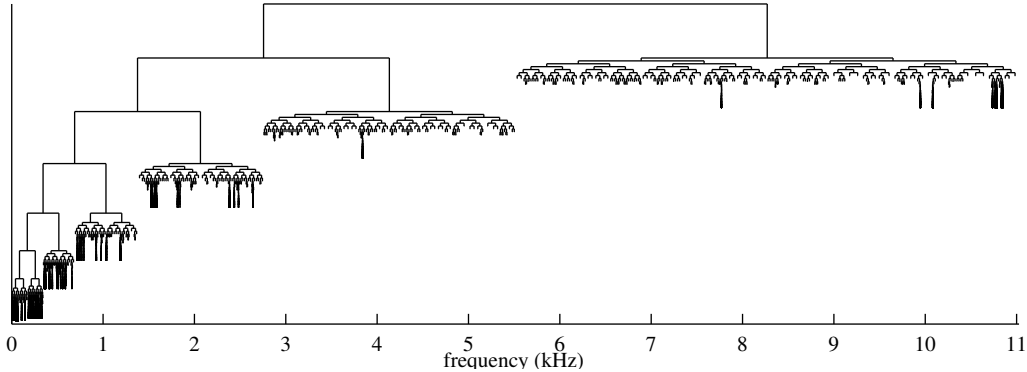


Figure 14: Tree representation of the oracle generic WP basis for real-valued masking of music mixtures. The height of each branch is proportional to the decrease in distortion \tilde{d} obtained by splitting.

determined by (26) for real-valued masking of single-channel speech or music mixtures. The trees displayed in these figures indicate which splits of the frequency axis actually improve the average oracle separation performance, with the amount of improvement being indicated by the height of the corresponding branches. One can see that in both cases the structure of the oracle generic WP basis is similar to that of a standard wavelet basis: recursively splitting the lower part of the frequency axis yields a greater decrease of the resulting distortion than splitting its upper part. When compared to the systematic use of an adapted oracle WP basis $\tilde{\lambda}_l$ for each example, using the same generic oracle WP basis $\tilde{\lambda}$ for all examples decreases the SDR by 0.5 dB for speech data and 1.4 dB for music data.

8 Example applications

After having studied each oracle estimator separately in the previous sections, we now provide a cross-comparison of several oracle estimators and some blind separation algorithms. Three distinct applications of oracle estimators are considered. 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 various applications on a set of examples.

8.1 Experimental setup

We considered four experimental configurations corresponding to four types of realistic mixtures, each characterized by a specific set of mixing filters and a type of audio sources: 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). In each configuration, we generated ten mixtures by applying the specified mixing filters to ten different sets of sources of the specified type.

We compared near-optimal estimators for frequency-domain beamforming (see section 4.3.4) and multichannel binary masking (see section 6.3.2) with two blind source separation algorithms belonging to the same two classes: the Frequency-Domain ICA (FDICA) algorithm presented in [16] 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 6.3.2).

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

Table 1: Comparison of oracle estimators and blind source separation algorithms on speech and music mixtures. Bounds indicate 95% confidence intervals. Note that 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-optimal beamforming	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

to determined and over-determined mixtures ($J \leq I$) and DUET to stereo mixtures ($I = 2$), but that near-optimal estimators can be applied in all configurations. For FDICA, near-optimal beamforming 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 6.3.2. 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 configuration, with 95% confidence bounds quantifying performance variability depending on the sources (mixing filters being fixed for each configuration). We discuss the main observations below.

8.2 Oracle vs. blind separation

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 beamforming 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 configuration. While the performance of near-optimal beamforming 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 assumption 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 beamforming 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.

8.3 Beamforming *vs.* time-frequency masking

Comparing the performance of the various near-optimal estimators, one can observe that near-optimal beamforming exhibits better performance (by at least 2 dB) than near-optimal binary masking, even for under-determined mixtures. This is likely to be because beamforming effectively combines all mixture channels in the separation function, despite its use of time-invariant filters. 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, beamforming is potentially the best class of separation algorithms for over-determined speech mixtures, where it outperforms local mixing inversion by about 6 dB. In all other configurations, local mixing inversion is preferable and its performance exceeds that of beamforming 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 [30] and convolutive mixtures [32], 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.

8.4 Towards objective 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.* [36, 18], but these are often specific to a particular objective function and not easily related to the achievable performance. For example, the measures in [36] 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.

9 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, 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.

9.1 Robustness of beamforming

Figure 15 displays the results for the time-domain beamforming 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.

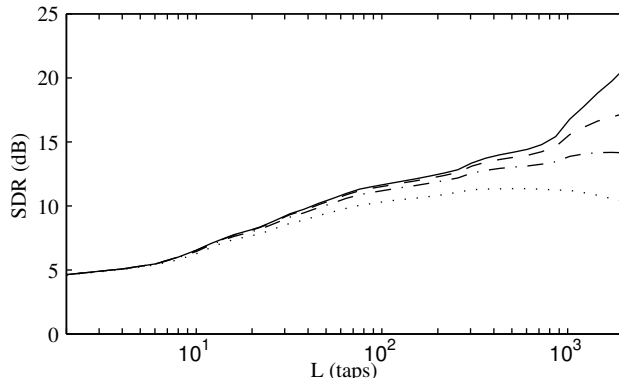


Figure 15: Average performance of the beamforming oracle on determined two-source mixtures with reverberation time $RT = 250$ ms as a function of the length L of the demixing filters. Each curve corresponds to a different hypothesized direction of source 1 for which the oracle demixing filters were learnt (plain: true direction, dashed: 2° error, dash-dotted: 4° error, dotted: 8° error).

Previously other authors [12] 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.

9.2 Robustness of local mixing inversion

Figure 16 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).

10 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 blind algorithms,

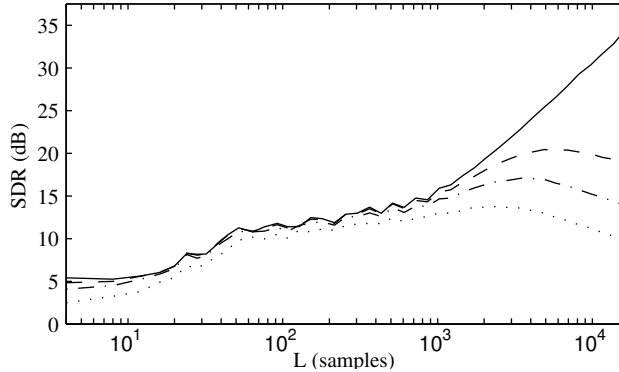


Figure 16: Average performance of near-optimal local mixing inversion with a free number of active sources $J'(n, f)$ for each (n, f) on determined two-source mixtures with reverberation time $RT = 250$ ms as a function of the STFT length L . Each curve corresponds to a different hypothesized direction of source 1 for which the oracle activity patterns were learnt (plain: true direction, dashed: 2° error, dash-dotted: 4° error, dotted: 8° error).

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 four particular classes of algorithms: beamforming, single-channel time-frequency masking, multichannel time-frequency masking and best basis 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 beamforming on determined mixtures by up to 20 dB compared to anechoic mixing, and that the best window length for single-channel time-frequency masking equals about 55 ms for speech and 190 ms for music. Secondly, we reported a few results markedly different from those of previous studies. For instance, we proved 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. We also showed that, even in case of source movements, beamforming filters of about a thousand taps remain preferable to shorter filters for the separation of reverberant mixtures. Thirdly, we were able to quantify the dependence of oracle beamforming filters on the source signals and the effect of frequency-domain implementation of beamforming, 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 beamforming and binary masking on determined and under-determined convolutive mixtures, by at least 4 dB and 8 dB respectively. 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. Best basis masking seems less promising, since oracle bases did not lead to substantially better performance than standard bases.

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, the oracle basis selection algorithm could be combined with an embedded blind mask estimation algorithm such as DUET, instead of using embedded oracle masks, to determine whether the use of the best basis significantly improves performance in a blind context. It would also 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 bases and oracle source patterns for time-frequency masking could give an insight into which objective functions would be suitable for the blind estimation of these quantities.

Acknowledgment

The authors wish to thank Scott Rickard and Nikolaos Mitianoudis for providing implementations of DUET and FDICA.

References

- [1] P. Smaragdis, “Blind separation of convolved mixtures in the frequency domain,” *Neurocomputing*, vol. 22, pp. 21–34, 1998.
- [2] J.-F. Cardoso, “Blind source separation : statistical principles,” *Proceedings of the IEEE*, vol. 9, no. 10, pp. 2009–2025, oct. 1998.
- [3] D.-T. Pham and J.-F. Cardoso, “Blind separation of instantaneous mixtures of non stationary sources,” *IEEE Trans. on Signal Processing*, vol. 49, no. 9, pp. 1837–1848, 2001.
- [4] 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.
- [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, 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.
- [7] Ö. Yilmaz 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.
- [8] N. Roman, D. 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.
- [9] 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.
- [10] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, “The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech,” *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 2, pp. 109–116, 2003.
- [11] A. Westner and 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.
- [12] 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.

- [13] 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.
- [14] K. E. Hild, D. Pinto, D. Erdogmus, and J. C. Principe, “Experimental upper bound for the performance of convolutive source separation methods,” *IEEE Trans. on Signal Processing*, vol. 54, no. 2, pp. 627–635, 2006.
- [15] K. Matsuoka and 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.
- [16] 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.
- [17] T. Takatani, T. Nishikawa, H. Saruwatari, and 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.
- [18] D. Schobben, K. Torkkola, and 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.
- [19] 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.
- [20] S. van de Par, A. Kohlrausch, G. Charestan, and 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.
- [21] U. P. Svensson and 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.
- [22] J. Nocedal and S. J. Wright, *Numerical optimization*. New York, NY: Springer, 1999.
- [23] D. W. Griffin and J. S. Lim, “Signal estimation from modified short-time Fourier transform,” *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.
- [24] 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.
- [25] J. P. Princen and A. B. Bradley, “Analysis/synthesis filter bank design based on time domain aliasing cancellation,” *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 34, no. 5, pp. 1153–1161, 1986.
- [26] S. Araki, S. Makino, H. Sawada, and 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.
- [27] R. Gribonval, “Piecewise linear source separation,” in *Proc. SPIE*, vol. 5207 Wavelets: Applications in Signal and Image Processing X, 2003, pp. 297–310.
- [28] 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.
- [29] A. Aïssa-El-Bey, K. Abed-Meraim, and 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.

- [30] P. Bofill and M. Zibulevsky, “Underdetermined blind source separation using sparse representations,” *Signal Processing*, vol. 81, pp. 2353–2362, 2001.
- [31] 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.
- [32] S. Winter, H. Sawada, and 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.
- [33] A. Nesbit, M. E. Davies, M. D. Plumbley, and M. B. Sandler, “Source extraction from two-channel mixtures by joint cosine packet analysis,” in *Proc. European Signal Processing Conf. (EUSIPCO)*, 2006, to appear.
- [34] R. R. Coifman and M. V. Wickerhauser, “Entropy-based algorithms for best basis selection,” *IEEE Trans. on Information Theory*, vol. 38, no. 2, pp. 713–718, 1992.
- [35] S. Mallat, *A Wavelet Tour of Signal Processing*. San Diego, CA: Academic Press, 1998.
- [36] 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.