

# Performance measurement in blind audio source separation

Emmanuel Vincent, Rémi Gribonval, Cédric Févotte

## ► To cite this version:

Emmanuel Vincent, Rémi Gribonval, Cédric Févotte. Performance measurement in blind audio source separation. IEEE Transactions on Audio, Speech and Language Processing, Institute of Electrical and Electronics Engineers, 2006, 14 (4), pp.1462–1469. <inria-00544230>

**HAL Id: inria-00544230**

**<https://hal.inria.fr/inria-00544230>**

Submitted on 7 Dec 2010

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Performance Measurement in Blind Audio Source Separation

Emmanuel Vincent\*, Rémi Gribonval and Cédric Févotte

**Abstract**—In this article, we discuss the evaluation of Blind Audio Source Separation (BASS) algorithms. Depending on the exact application, different distortions can be allowed between an estimated source and the wanted true source. We consider four different sets of such allowed distortions, from time-invariant gains to time-varying filters. In each case we decompose the estimated source into a true source part plus error terms corresponding to interferences, additive noise and algorithmic artifacts. Then we derive a global performance measure using an energy ratio, plus a separate performance measure for each error term. These measures are computed and discussed on the results of several BASS problems with various difficulty levels.

**Index Terms**—Audio source separation, performance, quality, evaluation, measure.

## I. INTRODUCTION

**B**LIND Audio Source Separation (BASS) has been a topic of intense work during the last years. Several successful models have emerged, such as Independent Component Analysis (ICA) [1], Sparse Decompositions (SD) [2] and Computational Auditory Scene Analysis (CASA) [3]. However, it is still hard to evaluate an algorithm or to compare several algorithms because of the lack of appropriate performance measures and common test sounds, even in the very simple case of linear instantaneous mixtures. In this article we design new numerical performance criteria that can help evaluate and compare algorithms when applied on usual BASS problems. Before we present these, let us first describe the problems considered and discuss the existing performance measures and their drawbacks.

### A. BASS general notations

The BASS problem arises when one or several microphones record a sound that is the mixture of sounds coming from several sources. For simplicity, we consider here only linear time-invariant mixing systems. If we denote by  $s_j(t)$  the signal emitted by the  $j$ -th source ( $1 \leq j \leq n$ ),  $x_i(t)$  the signal recorded by the  $i$ -th microphone ( $1 \leq i \leq m$ ) and  $a_{ij}(\tau)$  the (causal) source-to-microphone filters, we have  $x_i(t) = \sum_{j=1}^n \sum_{\tau=0}^{+\infty} a_{ij}(\tau) s_j(t - \tau) + n_i(t)$ , where  $n_i(t)$  is some

additive sensor noise. This  $m \times n$  mixture is expressed more conveniently using the matrix of filters formalism as

$$\mathbf{x} = \mathbf{A} \star \mathbf{s} + \mathbf{n}. \quad (1)$$

where  $\star$  denotes convolution. In the following, variables without time index will denote batch sequences, *e.g.*  $\mathbf{x} = [\mathbf{x}(0), \dots, \mathbf{x}(T-1)]$ . Bold letters will be used for multichannel variables, such as the vector of observations  $\mathbf{x}$ , the vector of sources  $\mathbf{s}$ , or the mixing system  $\mathbf{A}$ , and plain letters for monochannel variables, such as the  $j$ -th source  $s_j$ .

### B. BASS applications and difficulty levels

BASS covers many applications [4], and the criteria used to assess the performance of an algorithm depend on the application. Sometimes the goal is to extract source signals that are listened to, straight after separation or after some post-processing audio treatment. Sometimes, it is to retrieve source features and/or mixing parameters to describe complex audio scenes in a way related to human hearing. In this paper, we focus on the most common task addressed by BASS algorithms: Source Extraction.

Source Extraction consists in extracting from a mixture one or several mono source signals  $s_j$ . Examples include the denoising and dereverberation of speech for auditory prostheses and the extraction of interesting sounds in musical excerpts for electronic music creation. Without specific prior information about the sources  $\mathbf{s}$  or the mixing system  $\mathbf{A}$  this problem suffers from well-known theoretical indeterminacies [1], [5]. Generally the sources can only be recovered up to a permutation and arbitrary gains, but further indeterminacies may exist in convolutive mixtures (*e.g.* up to a filter).

Source Extraction can be addressed at various difficulty levels depending on the structure of the mixing system [6], [7]. A first difficulty criterion is the respective number of sources and sensors. In noiseless determined instantaneous mixtures (*i.e.* when  $m = n$ ) there exists a time-invariant linear demixing system  $\mathbf{W} = \mathbf{A}^{-1}$ . After  $\mathbf{W}$  has been estimated, the sources can be simply recovered as  $\mathbf{s} = \mathbf{W}\mathbf{x}$ . In noiseless under-determined mixtures (*i.e.* when  $m < n$ ), this is not possible anymore since the equation  $\mathbf{x} = \mathbf{A}\mathbf{s}$  has an affine set of solutions. This non trivial indeterminacy can be removed using knowledge about the sources, such as sparse priors [2]. A second difficulty criterion is the length of the mixing filters. Many algorithms for instantaneous noiseless determined mixtures provide near perfect results [1], [2], [8]. But convolutive mixtures still raise challenging theoretical issues such as the identifiability of the sources up to gain and technical difficulties like the estimation of long mixing filters in short duration mixtures.

Submitted to IEEE Transactions on Speech and Audio Processing on June 9th 2004. Revised on February 1st 2005. Accepted on May 3rd 2005.

Emmanuel Vincent is with Queen Mary, University of London, Electronic Engineering Department, Mile End Road, London E1 4NS (United Kingdom) (Phone: +44 20 7882 5528, Fax: +44 20 7882 7997, Email: emmanuel.vincent@elec.qmul.ac.uk). Rémi Gribonval is with IRISA, Campus de Beaulieu, F-35042 Rennes Cedex (France) (Phone: +33 2 9984 2506, Fax: +33 2 9984 7171, Email: remi.gribonval@irisa.fr). Cédric Févotte is with Cambridge University Engineering Department, Trumpington Street, Cambridge CB2 1PZ (United Kingdom) (Phone: +44 12 2376 5582, Fax: +44 12 2333 2662, Email: cf269@eng.cam.ac.uk)

### C. Existing performance measures and their limitations

Some performance measures for Source Extraction have already been defined in the literature. A first kind of measure assumes that the estimated sources  $\hat{s}$  have been recovered by applying a time-invariant linear demixing system  $\mathbf{W}$  to the observations  $\mathbf{x}$ . The global system  $\mathbf{B} = \mathbf{W} \star \mathbf{A}$  verifies  $\hat{s} = \mathbf{B} \star s$ . The quality of  $\hat{s}_j$  is then measured by the row Inter-Symbol Interference [7]

$$\text{ISI}_j := \frac{\sum_{j', \tau} |B_{jj'}(\tau)|^2 - \max_{j', \tau} |B_{jj'}(\tau)|^2}{\max_{j', \tau} |B_{jj'}(\tau)|^2}. \quad (2)$$

$\text{ISI}_j$  is always positive and equal to zero only when  $\hat{s}_j$  is equal to the true source  $s_{j'}$  up to a gain and a delay  $\tau$  with  $(j', \tau) = \arg \max_{j', \tau} |B_{jj'}(\tau)|^2$ . This criterion and other similar ones [9] are relevant, but cannot be applied to under-determined BASS problems because a perfect time-invariant demixing system  $\mathbf{W}$  does not exist generally. Moreover even in determined BASS it is possible to use other separation schemes than time-invariant linear demixing. A second kind of measure consists in comparing directly  $\hat{s}_j$  and  $s_j$ , paying attention to the indeterminacies of the task. The gain indeterminacy can be handled by comparing  $L_2$ -normalized versions of the sources with the relative square distance [6], [10], [2]

$$D := \min_{\epsilon = \pm 1} \left\| \frac{\hat{s}_j}{\|\hat{s}_j\|} - \epsilon \frac{s_j}{\|s_j\|} \right\|^2. \quad (3)$$

This measure is also relevant since it is always positive and equal to zero only when  $\hat{s}_j$  equals  $s_j$  up to a gain. However  $D$  takes at most the value  $D = 2$ , even in the worst case where the permutation indeterminacy has been badly solved and where  $\hat{s}_j$  equals another source  $s_{j'}$  orthogonal to  $s_j$ . One would then desire a distortion  $D = +\infty$ . More generally  $D$  evaluates bad results rather coarsely. For example  $\hat{s}_j = s_{j'}$  and  $\hat{s}_j = s_{j'}/\|s_{j'}\| + 0.02 s_j/\|s_j\|$  lead to similar measures  $D = 2$  and  $D \approx 1.96$  but are perceived quite differently.

These performance measures suffer from further limitations. Both consider only the case where  $\hat{s}_j$  has to be recovered up to a permutation and a gain. But in some applications it may be relevant to allow more or less distortions, not necessarily related to the theoretical indeterminacies of the problem. For example in hi-fi musical applications it may be important to recover the sources up to a simple gain since arbitrary filtering modifies the timbre of musical instruments. On the contrary in speech applications some filtering distortion may be allowed because low-pass filtered speech is generally still intelligible. Moreover both measures provide a single performance criterion containing all estimation errors. But in audio applications it is important to measure separately the amount of interferences from non-wanted sources, the amount or remaining sensor noise and the amount of “burbling” artifacts (also termed “musical noise”). Such artifacts are often considered as a more annoying kind of error than interferences, that are themselves more annoying than sensor noise. Many separation methods for under-determined BASS problems produce few interferences but many artifacts [11], [12], [13], and this cannot be described by a single criterion.

### D. Overview of our proposals

The goal of this article is to design new performance criteria that can be applied on all usual BASS problems and overcome the limitations above. The only assumptions we make is that

- the true source signals and noise signals (if any) are known,
- the user chooses a family of allowed distortions  $\mathcal{F}$  according to the application (but independently of the kind of mixture or the algorithm used).

The mixing system and the demixing technique do not need to be known.

Separate performance measures are computed for each estimated source  $\hat{s}_j$  by comparing it to a given true source  $s_j$ . Note that the measures do not take into account the permutation indeterminacy of BASS. If necessary,  $\hat{s}_j$  may be compared with all the sources  $(s_{j'})_{1 \leq j' \leq n}$  and the “true source” may be selected as the one that gives the best results.

The computation of the criteria involves two successive steps. In a first step, we decompose  $\hat{s}_j$  as

$$\hat{s}_j = s_{\text{target}} + e_{\text{interf}} + e_{\text{noise}} + e_{\text{artif}}, \quad (4)$$

where  $s_{\text{target}} = f(s_j)$  is a version of  $s_j$  modified by an allowed distortion  $f \in \mathcal{F}$ , and where  $e_{\text{interf}}$ ,  $e_{\text{noise}}$  and  $e_{\text{artif}}$  are respectively the interferences, noise and artifacts error terms. These four terms should represent the part of  $\hat{s}_j$  perceived as coming from the wanted source  $s_j$ , from other unwanted sources  $(s_{j'})_{j' \neq j}$ , from sensor noises  $(n_i)_{1 \leq i \leq m}$ , and from other causes (like forbidden distortions of the sources and/or “burbling” artifacts). In a second step, we compute energy ratios to evaluate the relative amount of each of these four terms either on the whole signal duration or on local frames.

### E. Structure of the article

The rest of the article has the following structure. In Section II we show how to decompose  $\hat{s}_j$  and compute the performance measures when  $\mathcal{F}$  is the set of time-invariant gains distortions (this covers our preliminary proposals introduced in [14]). In Section III we extend these results to the case where  $\mathcal{F}$  contains time-varying and/or filtering distortions. In Section IV we test the measures on several BASS problems. In Section V we discuss their relevance for algorithm evaluation and comment their correlation with subjective performance on informal listening tests. Finally we conclude in Section VI by pointing out further perspectives about BASS evaluation and introducing our on-line evaluation database *BASS-dB* [15].

## II. PERFORMANCE CRITERIA FOR TIME-INVARIANT GAINS ALLOWED DISTORTIONS

We propose in this section performance criteria for the most usual case when the only allowed distortions on  $\hat{s}_j$  are time-invariant gains. We first show how to decompose  $\hat{s}_j$  into four terms as in (4) and then we define relevant energy ratios between these terms.

Let us denote in the following  $\langle a, b \rangle := \sum_{t=0}^{T-1} a(t)\bar{b}(t)$  the inner product between two possibly complex-valued<sup>1</sup> signals  $a$  and  $b$  of length  $T$ , where  $\bar{b}$  is the complex conjugate of  $b$ , and  $\|a\|^2 := \langle a, a \rangle$  the energy of  $a$ .

#### A. Estimated source decomposition by orthogonal projections

When  $\mathbf{A}$  is a time-invariant instantaneous matrix and when the mixture is separated by applying a time-invariant instantaneous matrix  $\mathbf{W}$ ,  $\hat{s}_j$  can be decomposed as

$$\hat{s}_j = (\mathbf{WA})_{jj}s_j + \sum_{j' \neq j} (\mathbf{WA})_{jj'}s_{j'} + \sum_{i=1}^m W_{ji}n_i. \quad (5)$$

Since  $(\mathbf{WA})_{jj}$  is a time-invariant gain, it seems natural to identify the three terms of this sum with  $s_{\text{target}}$ ,  $e_{\text{interf}}$  and  $e_{\text{noise}}$  respectively ( $e_{\text{artif}} = 0$  here). However (5) cannot be used as a definition of  $s_{\text{target}}$ ,  $e_{\text{interf}}$ ,  $e_{\text{noise}}$  and  $e_{\text{artif}}$  since the mixing and demixing systems are unknown. Also the two first terms of (5) may not be perceived as separate sound objects when a nonwanted source  $s_{j'}$  is highly correlated with the wanted source  $s_j$ .

Instead, the decomposition we propose is based on orthogonal projections. Let us denote  $\Pi\{y_1, \dots, y_k\}$  the orthogonal projector onto the subspace spanned by the vectors  $y_1, \dots, y_k$ . The projector is a  $T \times T$  matrix, where  $T$  is the length of these vectors. We consider the three orthogonal projectors

$$P_{s_j} := \Pi\{s_j\}, \quad (6)$$

$$P_{\mathbf{s}} := \Pi\{(s_{j'})_{1 \leq j' \leq n}\}, \quad (7)$$

$$P_{\mathbf{s}, \mathbf{n}} := \Pi\{(s_{j'})_{1 \leq j' \leq n}, (n_i)_{1 \leq i \leq m}\}. \quad (8)$$

And we decompose  $\hat{s}_j$  as the sum of the four terms

$$s_{\text{target}} := P_{s_j}\hat{s}_j, \quad (9)$$

$$e_{\text{interf}} := P_{\mathbf{s}}\hat{s}_j - P_{s_j}\hat{s}_j, \quad (10)$$

$$e_{\text{noise}} := P_{\mathbf{s}, \mathbf{n}}\hat{s}_j - P_{\mathbf{s}}\hat{s}_j, \quad (11)$$

$$e_{\text{artif}} := \hat{s}_j - P_{\mathbf{s}, \mathbf{n}}\hat{s}_j. \quad (12)$$

The computation of  $s_{\text{target}}$  is straightforward since it involves only a simple inner product:  $s_{\text{target}} = \langle \hat{s}_j, s_j \rangle s_j / \|s_j\|^2$ . The computation of  $e_{\text{interf}}$  is a bit more complex. If the sources are mutually orthogonal, then  $e_{\text{interf}} = \sum_{j' \neq j} \langle \hat{s}_j, s_{j'} \rangle s_{j'} / \|s_{j'}\|^2$ . Otherwise, if we use a vector  $\mathbf{c}$  of coefficients such that  $P_{\mathbf{s}}\hat{s}_j = \sum_{j'=1}^n \bar{c}_{j'} s_{j'} = \mathbf{c}^H \mathbf{s}$  (where  $(\cdot)^H$  denotes Hermitian transposition), then  $\mathbf{c} = \mathbf{R}_{\mathbf{ss}}^{-1} [\langle \hat{s}_j, s_1 \rangle, \dots, \langle \hat{s}_j, s_n \rangle]^H$ , where  $\mathbf{R}_{\mathbf{ss}}$  is the Gram matrix of the sources defined by  $(\mathbf{R}_{\mathbf{ss}})_{jj'} = \langle s_j, s_{j'} \rangle$ . The computation of  $P_{\mathbf{s}, \mathbf{n}}$  proceeds in a similar fashion, however most of the time we can make the assumption that the noise signals are mutually orthogonal and orthogonal to each source, so that  $P_{\mathbf{s}, \mathbf{n}}\hat{s}_j \approx P_{\mathbf{s}}\hat{s}_j + \sum_{i=1}^m \langle \hat{s}_j, n_i \rangle n_i / \|n_i\|^2$ .

<sup>1</sup> Audio signals are real-valued, but it is costless to express our performance criteria in the slightly more general complex setting which might be useful for other types of signals.

#### B. From estimated source decomposition to global performance measures

Starting from the decomposition of  $\hat{s}_j$  in (6-12), we now define numerical performance criteria by computing energy ratios expressed in decibels (dB). We define the Source to Distortion Ratio

$$\text{SDR} := 10 \log_{10} \frac{\|s_{\text{target}}\|^2}{\|e_{\text{interf}} + e_{\text{noise}} + e_{\text{artif}}\|^2}, \quad (13)$$

the Source to Interferences Ratio

$$\text{SIR} := 10 \log_{10} \frac{\|s_{\text{target}}\|^2}{\|e_{\text{interf}}\|^2}, \quad (14)$$

the Sources to Noise Ratio

$$\text{SNR} := 10 \log_{10} \frac{\|s_{\text{target}} + e_{\text{interf}}\|^2}{\|e_{\text{noise}}\|^2}, \quad (15)$$

and the Sources to Artifacts Ratio

$$\text{SAR} := 10 \log_{10} \frac{\|s_{\text{target}} + e_{\text{interf}} + e_{\text{noise}}\|^2}{\|e_{\text{artif}}\|^2}. \quad (16)$$

These four measures are inspired by the usual definition of the SNR, with a few modifications. For instance the definition of the SNR involving the term  $s_{\text{target}} + e_{\text{interf}}$  at the numerator aims at making it independent of the SIR. Indeed consider the case of an instantaneous noisy  $2 \times 2$  mixture where  $\hat{s}_1 = \epsilon s_1 + s_2 + e_{\text{noise}}$  with  $\|\epsilon s_1\| \ll \|s_2\|$  and  $\|e_{\text{noise}}\| \approx \|\epsilon s_1\|$ . Then  $\hat{s}_1$  is perceived as dominated by the interfering signal, with the noise energy making an insignificant contribution. This is consistent with  $\text{SIR} \approx -\infty$  and  $\text{SNR} \approx +\infty$  using our definitions. A SNR defined by  $10 \log_{10} (\|s_{\text{target}}\|^2 / \|e_{\text{noise}}\|^2)$  would give  $\text{SNR} \approx 0$  instead. Similarly, the SAR is independent of the SIR and the SNR since the numerator in (16) includes the interferences and noise terms as well.

Note that the numerical precision of the measures is lower for high performance values than for low ones. For example a high SDR means that the denominator in (13) is very small, so that small constant-amplitude errors in  $s_{\text{target}}$  (due to signal quantization) result in large SDR deviations. In particular, when the signals correspond to sound files, the precision of the results depends on the number of bits per sample.

#### C. Local performance measures

When the powers of  $s_{\text{target}}$ ,  $e_{\text{interf}}$ ,  $e_{\text{noise}}$  and  $e_{\text{artif}}$  vary across time, the perceived separation quality also varies accordingly. We take this into account by defining local numerical performance measures in the following way.

First we compute  $s_{\text{target}}$ ,  $e_{\text{interf}}$ ,  $e_{\text{noise}}$  and  $e_{\text{artif}}$  as in (6-12). Then, denoting  $w$  a finite length centered window, we compute the windowed signals  $s_{\text{target}}^r$ ,  $e_{\text{interf}}^r$ ,  $e_{\text{noise}}^r$  and  $e_{\text{artif}}^r$  centered in  $r$ , where  $s_{\text{target}}^r(t) = w(t-r)s_{\text{target}}(t)$  and so on. Finally, for each  $r$  we compute the local measures  $\text{SDR}^r$ ,  $\text{SIR}^r$ ,  $\text{SNR}^r$  and  $\text{SAR}^r$  as in (13-16) but replacing the original terms by the windowed terms centered in  $r$ .

$\text{SDR}^r$ ,  $\text{SIR}^r$ ,  $\text{SNR}^r$  and  $\text{SAR}^r$  thus measure the separation performance on the time frame centered in  $r$ . All these values

can be visualized more globally by plotting them against  $r$  or by summarizing them into cumulative histograms [13]. Global performance measures can also be defined in the spirit of segmental SNR [16]. The shape of the window  $w$  has not much importance generally, only its duration is relevant. Thus a rectangular window may be used for simplicity.

#### D. Comparison with existing performance measures

The new performance measures solve most of the problems encountered with existing measures discussed in Section I-C.

First the computation does not rely on the assumption that a particular type of demixing system or algorithm is used. The only assumption is that  $\hat{s}_j$  has to be recovered up to a time-invariant gain. Measures for other allowed distortions are proposed in Section III.

Secondly the SDR has better properties than D. Simple calculus shows that both measures are identical up to a one-to-one mapping  $10^{-\text{SDR}/10} = D(4-D)/(2-D)^2$ . However, contrary to  $-10 \log_{10} D$ , the SDR is not lower-bounded:  $\text{SDR} = -\infty$  when  $s_{\text{target}} = 0$  and evaluation of bad results is less coarse [14].

Thirdly four criteria are proposed instead of a single one. The SIR, SNR and SAR allow to distinguish between estimation errors that are mostly dominated by interferences, noise or artifacts. This is verified on test mixtures in Section IV.

### III. PERFORMANCE CRITERIA FOR OTHER ALLOWED DISTORTIONS

#### A. Which equations have to be modified ?

After having defined performance measures for time-invariant gains allowed distortions, we consider now similar measures for other allowed distortions. Much of the work done in the previous section is still relevant here and only the definitions of the orthogonal projectors in (6-8) have to be modified.

Indeed the two steps consisting in decomposing  $\hat{s}_j$  in four terms and in computing energy ratios between these terms do not depend on each other. Since the kind of allowed distortion is not used in the second step, the performance measures are always defined by (13-16), whatever the allowed distortion is.

Moreover the decomposition of  $\hat{s}_j$  can also often be defined by (9-12), but using other orthogonal projectors depending on the allowed distortions. In the following we present successively the projectors used to decompose  $\hat{s}_j$  when filtering and/or time-varying distortions are allowed.

#### B. Time-invariant filters allowed distortions

When time-invariant filters are allowed,  $s_{\text{target}}$  is not a scaled version of  $s_j$  anymore, but a filtered version expressed as  $s_{\text{target}}(t) = \sum_{\tau=0}^{L-1} h(\tau) \times s_j(t-\tau)$ . If we express this in terms of subspaces,  $s_{\text{target}}$  does not generally belong to the subspace spanned by  $s_j$  but to the subspace spanned by delayed versions of  $s_j$ . So we can define  $s_{\text{target}}$  by projecting  $\hat{s}_j$  on this new subspace.

We denote by  $s_j^\tau$  and  $n_i^\tau$  the source signal  $s_j$  and the noise signal  $n_i$  delayed by  $\tau$ , so that  $s_j^\tau(t) = s_j(t-\tau)$  and  $n_i^\tau(t) =$

$n_i(t-\tau)$ . To avoid multiple definitions due to boundary effects, we consider the support of all signals to be  $[0, T+L-2]$  where  $[0, T-1]$  is the original support of the signals and  $L-1$  is the maximum delay allowed. We define the decomposition using the three projectors

$$P_{s_j} := \Pi\{(s_j^\tau)_{0 \leq \tau \leq L-1}\}, \quad (17)$$

$$P_s := \Pi\{(s_{j'}^\tau)_{1 \leq j' \leq n, 0 \leq \tau \leq L-1}\}, \quad (18)$$

$$P_{s,n} := \Pi\{[(s_{j'}^\tau)_{1 \leq j' \leq n}, (n_i^\tau)_{1 \leq i \leq m}]_{0 \leq \tau \leq L-1}\}, \quad (19)$$

The computation of the projections again involves inversion of Gram matrices. The Gram matrix corresponding to  $P_{s_j}$  is the empirical auto-correlation matrix of  $s_j$  defined by  $(\mathbf{R}_{s_j s_j})_{\tau\tau'} = \langle s_j^\tau, s_j^{\tau'} \rangle$ . The Gram matrix associated with  $P_s$  has a symmetric block-Toeplitz structure, where the blocks on the  $\tau$ -th diagonal are the empirical auto-correlation matrix of the sources at lag  $\tau$  defined by  $(\mathbf{R}_{ss}(\tau))_{jj'} = \langle s_j, s_{j'}^\tau \rangle$ .

#### C. Time-varying gains allowed distortions

When time-varying gains distortions are allowed,  $s_{\text{target}}$  is equal to  $s_j$  multiplied by a slowly time-varying gain. We parameterize this gain as  $g(t) = \sum_{u=0}^{U-1} \alpha_u v(t-uT')$ , where  $v$  is a positive kernel (i.e. a window) of length  $L'$  and  $T'$  the hopsize in samples between two successive “breakpoints”. When  $v$  is a rectangular window and  $L' = T'$ , this parameterization yields piecewise constant gains with breakpoints at  $uT'$ , but choosing a smoother kernel makes it possible to get more smoothly varying gains. This gives  $s_{\text{target}}(t) = g(t)s_j(t) = \sum_{u=0}^{U-1} \alpha_u \times v(t-uT')s_j(t)$ . Thus  $s_{\text{target}}$  belongs to the subspace spanned by versions of  $s_j$  windowed by the kernel  $v$ . Note that this use of windowed source signals has no link with the computation of local performance measures from windowed decomposed signals in Section II-C. We emphasize again that the decomposition of  $\hat{s}_j$  and the computation of energy ratios are separate steps.

We define the windowed source signals  $(s_j^u)_{0 \leq u \leq U-1}$  and the windowed noise signals  $(n_i^u)_{0 \leq u \leq U-1}$  of support  $[0, T-1]$  by  $s_j^u(t) = v(t-uT')s_j(t)$  and  $n_i^u(t) = v(t-uT')n_i(t)$ . The projectors for decomposition are given by

$$P_{s_j} := \Pi\{(s_j^u)_{0 \leq u \leq U-1}\}, \quad (20)$$

$$P_s := \Pi\{(s_{j'}^u)_{1 \leq j' \leq n, 0 \leq u \leq U-1}\}, \quad (21)$$

$$P_{s,n} := \Pi\{[(s_{j'}^u)_{1 \leq j' \leq n}, (n_i^u)_{1 \leq i \leq m}]_{0 \leq u \leq U-1}\}. \quad (22)$$

In order to guarantee the natural property  $P_{s_j}s_j = s_j$  (i.e.  $\text{SDR} = +\infty$  as expected in the particular case where  $\hat{s}_j = s_j$ ), we enforce the condition that  $\sum_{u=0}^{U-1} v(t-uT')$  is a constant for all  $t$ . When this is verified,  $s_j$  belongs to the subspace spanned by  $(s_j^u)_{0 \leq u \leq U-1}$  because  $s_j = \sum_{u=0}^{U-1} s_j^u / \sum_{u=0}^{U-1} v(t-uT')$ . This condition always holds true when  $T' = 1$ , but other values of  $T'$  may work depending on the kernel  $v$ . For example if  $v$  is a triangular window and  $L'$  is a multiple of 2, then  $T' = L'/2$  also works.

#### D. Time-varying filters allowed distortions

Finally, when time-varying filters distortions are allowed, the decomposition of  $\hat{s}_j$  is made by combining the ideas of

the two previous subsections. The estimated source  $s_{\text{target}}$  is expressed as a version of  $s_j$  “convolved” by a slowly time-varying filter. Using the notations of the previous subsections, this results in  $s_{\text{target}}(t) = \sum_{\tau=0}^{L-1} h(\tau, t) s_j(t - \tau) = \sum_{\tau=0}^{L-1} \sum_{u=0}^{U-1} \alpha_{\tau u} \times v(t - uT') s_j(t - \tau)$ . Thus  $s_{\text{target}}$  belongs to the subspace spanned by delayed versions of  $s_j$  windowed by the kernel  $v$ .

We compute the windowed delayed source signals  $(s_j^{\tau u})_{0 \leq \tau \leq L-1, 0 \leq u \leq U-1}$  and the windowed delayed noise signals  $(n_i^{\tau u})_{0 \leq \tau \leq L-1, 0 \leq u \leq U-1}$  of support  $[0, T + L - 2]$  by windowing delayed signals:  $s_j^{\tau u}(t) = v(t - uT') s_j^{\tau}(t)$  and  $n_i^{\tau u}(t) = v(t - uT') n_i^{\tau}(t)$ . Note that the reverse order computation (passing windowed signals through delay lines) is not equivalent and results in other projections generally. We define the decomposition by the projectors

$$P_{s_j} := \Pi\{(s_j^{\tau u})_{0 \leq \tau \leq L-1, 0 \leq u \leq U-1}\}, \quad (23)$$

$$P_{\mathbf{s}} := \Pi\{(s_{j'}^{\tau u})_{1 \leq j' \leq n, 0 \leq \tau \leq L-1, 0 \leq u \leq U-1}\}, \quad (24)$$

$$P_{\mathbf{s}, \mathbf{n}} := \Pi\{(s_{j'}^{\tau u})_{1 \leq j' \leq n}, (n_i^{\tau u})_{1 \leq i \leq m}\}_{0 \leq \tau \leq L-1, 0 \leq u \leq U-1}\}. \quad (25)$$

#### IV. EVALUATION EXAMPLES

In order to assess the relevance of our performance measures, we made tests on a few usual BASS problems. The separated sources were either simulated from a known decomposition or extracted from the mixtures with existing BASS algorithms.

In this section we present the results of performance measurement on three noiseless mixtures of three musical sources. The sources are 16 bits sound files of 2.4 s, sampled at 8 KHz ( $T = 19200$ ) and normalized.  $s_1$  is cello,  $s_2$  drums and  $s_3$  piano. The three mixtures are 16 bits sound files containing an instantaneous  $2 \times 2$  mixture, a convolutive  $2 \times 2$  mixture and an instantaneous  $2 \times 3$  mixture. These mixtures were chosen because they have very different difficulty levels and they act as typical mixtures within the large amount of usual audio mixtures. We also chose some typical existing algorithms to separate these mixtures to show that the performance measures are relevant on “real life” data.

For each mixture and each algorithm the (non quantized) estimated sources are decomposed using different allowed distortions and the performance measures are computed. The results are summarized in Tables II, III and IV. The different kinds of allowed distortions and corresponding decompositions are denoted TI Gain, TI Filt, TV Gain and TV Filt respectively. The values of the decomposition parameters are listed in Table I. Their choice is discussed in Section V based on informal listening tests.

The sound files corresponding to these examples are available for listening on <http://www.irisa.fr/metiss/demos/bssperf/>. This demo web page provides the sound files of the mixture  $\mathbf{x}$ , the sources  $\mathbf{s}$ , the first estimated source  $\hat{s}_1$ , and also the sound files of  $s_{\text{target}}$ ,  $e_{\text{interf}}$ ,  $e_{\text{noise}}$  and  $e_{\text{artif}}$  from the decomposition of  $\hat{s}_1$ . Sound files of  $\hat{s}_2$ ,  $\hat{s}_3$  and their decompositions are not provided for the sake of legibility. We emphasize that listening to these sounds and comparing with the related performance figures is the best way to evaluate the meaningfulness of our proposals.

TABLE I  
PARAMETER VALUES USED FOR DECOMPOSITION (8 KHz SAMPLE RATE).

| Allowed distortion | Delays<br>$L$ | Time frames |      |      |
|--------------------|---------------|-------------|------|------|
|                    |               | $v$         | $L'$ | $T'$ |
| TI Gain            | N/A           | N/A         |      |      |
| TI Filt            | 256           | N/A         |      |      |
| TV Filt            | 64            | Rect        | 1600 | 1600 |

##### A. Instantaneous $2 \times 2$ mixture

Our first example is a stereo instantaneous mixture of  $s_1$  and  $s_2$ , obtained with the mixing matrix

$$\mathbf{A} = \begin{bmatrix} 0.5 & 1 \\ 1 & 0.5 \end{bmatrix}.$$

We solve this problem by estimating a demixing matrix with two different ICA methods : by using non Gaussian distributions and mutual independence of the sources with JADE [1], and by finding zones in the time-frequency plane where only one source is present with TFBSS [8] (used with 64 time frames and 1024 frequency bins as input parameters). The performance measures are shown in Table II for TI Gain decompositions. Results with other decompositions differ from less than 2 dB. Since the global mixing-unmixing system  $\mathbf{WA}$  is known, we also compute for comparison the System SIR (SSIR) which is the power ratio between the two first terms of (5).

As expected with sources estimated by time-invariant linear demixing, no artifacts come into play:  $\text{SAR} \approx +\infty$  up to numerical precision. The estimation error is dominated by interferences and  $\text{SDR} \approx \text{SIR}$ . Moreover, the SIR is higher for TFBSS than for JADE. This result is corroborated by the fact that the demixing matrix estimated with TFBSS is closer to the true demixing matrix than with JADE. Also, as expected the SSIR is very close to the SIR, because the correlation between the sources  $\langle s_1, s_2 \rangle = -0.0055$  is small.

TABLE II  
EVALUATION OF AN INSTANTANEOUS  $2 \times 2$  MIXTURE.

| Method | Allowed distortion | SDR (dB)    |             | SIR (dB)    |             | SAR (dB)    |             | SSIR (dB)   |             |
|--------|--------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
|        |                    | $\hat{s}_1$ | $\hat{s}_2$ | $\hat{s}_1$ | $\hat{s}_2$ | $\hat{s}_1$ | $\hat{s}_2$ | $\hat{s}_1$ | $\hat{s}_2$ |
| JADE   | TI Gain            | 26          | 25          | 26          | 25          | 72          | 73          | 26          | 25          |
| TFBSS  | TI Gain            | 37          | 34          | 37          | 34          | 73          | 73          | 37          | 34          |

##### B. Convolutive $2 \times 2$ mixture

Our second example is a convolutive mixture of  $s_1$  and  $s_2$  made with 256 tap filters. The problem is solved by a frequential domain ICA method using 256 sub-bands and separating the mixture with JADE [1] in each sub-band. The usual “permutation problem” [5] is encountered when building estimated sources from extraction results in each sub-band. We test two methods to address this problem: the method outlined in [5] and an oracle (*i.e.* choice of the optimal permutations knowing the true sources). The

corresponding algorithms are named Frequential ICA (FICA) or Oracle Frequential ICA (OFICA). Both methods do not aim at recovering the sources  $s_1$  and  $s_2$  but their images on the first channel  $a_{11} \star s_1$  and  $a_{12} \star s_2$ . Thus the estimated sources may be at best filtered versions of the true sources. The performance measures are shown in Table III for three different decompositions. We compute again the SSIR from (5), with **WA** now containing filters instead of gains.

Different conclusions arise depending on the decomposition. The TI Gain decomposition results in low SDRs for both methods and  $\text{SDR} \approx \text{SAR}$ . Many artifacts arise due to forbidden (filter) distortions of the sources. On the contrary the TI Filt decomposition outputs a high SDR for OFICA and a medium SDR for FICA with  $\text{SDR} \approx \text{SIR}$ . Artifacts are smaller because filter distortions are allowed, and interferences are larger for FICA than for OFICA because the use of oracle information in OFICA prevents bad pairing of sub-bands. The TV Filt decomposition leads to intermediate results. Short (64-tap) filter distortions are allowed, but longer filter distortions are not, thus some of the filter distortions on the estimated sources are counted towards artifacts. Again the SSIR is very close the SIR computed using the TI Filt decomposition, because the correlation between delayed versions of different sources is also small.

It is also interesting to study the evolution of the performance measures across time. For example Fig. 1 plots the local SIR for  $\hat{s}_1$  (estimated with FICA) using TI Filt. We see that the actual performance measure varies a lot, which cannot be explained by a single global SIR.

TABLE III  
EVALUATION OF A CONVOLUTIVE  $2 \times 2$  MIXTURE.

| Method | Allowed distortion | SDR (dB)    |             | SIR (dB)    |             | SAR (dB)    |             | SSIR (dB)   |             |
|--------|--------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
|        |                    | $\hat{s}_1$ | $\hat{s}_2$ | $\hat{s}_1$ | $\hat{s}_2$ | $\hat{s}_1$ | $\hat{s}_2$ | $\hat{s}_1$ | $\hat{s}_2$ |
| FICA   | TI Gain            | -11         | -16         | 7           | 12          | -10         | -16         |             |             |
|        | TI Filt            | 6           | 14          | 6           | 17          | 19          | 17          | 6           | 16          |
|        | TV Filt            | 3           | -2          | 8           | 9           | 5           | -1          |             |             |
| OFICA  | TI Gain            | -10         | -13         | 15          | 22          | -10         | -13         |             |             |
|        | TI Filt            | 10          | 16          | 10          | 18          | 19          | 19          | 10          | 17          |
|        | TV Filt            | 5           | -2          | 15          | 10          | 5           | -1          |             |             |

### C. Instantaneous $2 \times 3$ mixture

Our third example is an instantaneous mixture of  $s_1$ ,  $s_2$  and  $s_3$  computed with the mixing matrix

$$\mathbf{A} \approx \begin{bmatrix} 0.92 & 1.40 & 1.05 \\ 0.36 & 1.30 & 2.36 \end{bmatrix}.$$

To solve this problem, we represent the two mixture channels in a domain where the sources exhibit a sparse behavior, and then the mixing matrix and the sources are estimated by (nonlinear) clustering of the ratios of the representation coefficients between the channels. Two algorithms are tested: a clustering of the Short Time Fourier Transform (STFT) called DUET [12] (using a 256 sample Hanning window and a 192 sample hopsize for STFT computation) and a Matching

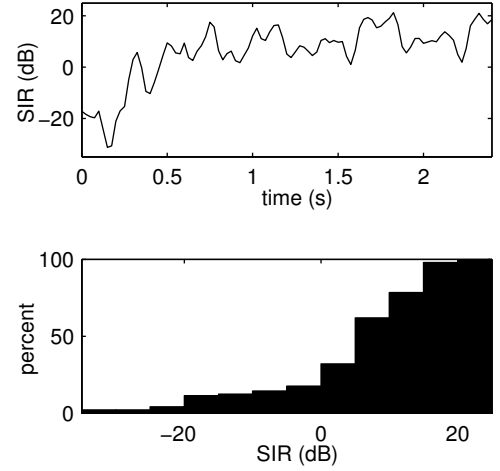


Fig. 1. Local SIR for  $\hat{s}_1$  estimated by FICA and TI Filt decomposition in the  $2 \times 2$  convolutive mixture. Hanning windows of length 100 ms and overlapping 75 ms are used. The SIR is plotted against time in the upper plot and summarized in a cumulative histogram in the lower plot.

Pursuit Clustering (MPC) [11] (using 10 000 Gabor atoms with truncated Gaussian envelope). The performance measures are shown in Table IV for two different decompositions. Unlike in the previous experiments, it does not seem possible to display any sort of “System SIR” for the MPC algorithm, since the result of the separation is not a linear function of the input sources. In a sense, this perfectly illustrates a situation where it is necessary to have at hand performance measures such as the ones we define, that is to say measures which do not rely on a particular type of separation algorithm, but simply try to compare the estimated signals with the target ones.

This time the choice of the decomposition has less influence on the results. Both methods estimate the sources essentially without filter distortions but with “burbling” artifacts due to source overlap in the representation domain. Thus SDRs are low for both methods and  $\text{SDR} \approx \text{SAR}$ . Note that MPC leads to better performance than DUET, particularly for  $\hat{s}_3$ . Indeed the use of an overcomplete dictionary in MPC makes the source representations sparser and limits source overlap.

TABLE IV  
EVALUATION OF AN INSTANTANEOUS  $2 \times 3$  MIXTURE.

| Method | Allowed distortion | SDR (dB)    |             |             | SIR (dB)    |             |             | SAR (dB)    |             |             |
|--------|--------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
|        |                    | $\hat{s}_1$ | $\hat{s}_2$ | $\hat{s}_3$ | $\hat{s}_1$ | $\hat{s}_2$ | $\hat{s}_3$ | $\hat{s}_1$ | $\hat{s}_2$ | $\hat{s}_3$ |
| STFTC  | TI Gain            | 2           | 4           | 4           | 15          | 9           | 29          | 3           | 7           | 4           |
|        | TI Filt            | 5           | 4           | 6           | 14          | 7           | 19          | 6           | 8           | 6           |
|        | TV Filt            | 6           | 5           | 8           | 11          | 7           | 15          | 8           | 11          | 9           |
| MPC    | TI Gain            | 4           | 8           | 15          | 19          | 27          | 31          | 4           | 8           | 15          |
|        | TI Filt            | 5           | 9           | 16          | 13          | 17          | 27          | 5           | 9           | 16          |
|        | TV Filt            | 6           | 9           | 16          | 14          | 14          | 23          | 7           | 11          | 17          |

## V. DISCUSSION

Before we conclude, let us summarize in this section the results of the evaluation examples. We discuss the relevance of

the proposed performance measures for algorithm evaluation and comparison. Then we describe how they could possibly be modified to explain subjective auditory assessments.

#### A. Relevance for algorithm evaluation and comparison

The main result of the previous section is that the SDR, SIR, SNR and SAR were found to be relevant for the evaluation of an algorithm and the comparison of several algorithms. Indeed, given a family of allowed distortions, the SIR and SAR were shown to be valid performance measures regarding two separate goals: rejection of the interferences and absence of forbidden distortions and “burbling” artifacts. Other experiments proved that the SNR was also valid for a third goal: rejection of the sensor noise. Finally the SDR was shown to be valid as a global performance measure in case these three goals are equally important.

Another important result is that the measures were found to depend a lot on the number of delays and time frames chosen for decomposition. Experimentally the more distortions are allowed the higher the SDRs are. More rigorously, when  $\mathcal{F}$  and  $\mathcal{F}'$  are two families with  $\mathcal{F} \subset \mathcal{F}'$ , the SDR of a given estimated source  $\hat{s}_j$  is higher allowing distortions in  $\mathcal{F}'$  than in  $\mathcal{F}$ . Indeed the projection subspaces verify  $\{f(s_j), f \in \mathcal{F}\} \subset \{f(s_j), f \in \mathcal{F}'\}$  and thus  $\|\hat{s}_j - P_{s_j}\hat{s}_j\|$  is smaller allowing distortions in  $\mathcal{F}'$  than in  $\mathcal{F}$ .

This confirms our main postulate that the evaluation of the performance of an algorithm only makes sense given a family of allowed distortions. This can be seen as a nice property if the distortions allowed for the desired application correspond to one of the families presented in sections II and III with a precisely known number of frequency sub-bands and time frames. But this is annoying when one has no idea about which distortions to allow. In that case, we cannot “recommend” a family of allowed distortions more than another one: the “best” choice really depends on the application.

Finally an interesting result is that in the previous section the results of algorithm classification according to mean SDR were always the same whatever decomposition was used. We make the hypothesis that this is not a coincidence and that the classification order is rather independent of the family of allowed distortions. Of course this hypothesis is based on very few experiments actually, but we think it would be interesting to validate or infirm it using more data. If it was true then algorithm classification would be greatly simplified. Indeed it would be unnecessary to test many families of allowed distortions before providing a global result: a single one would suffice.

#### B. Relevance towards subjective performance measures

Another interesting question is to study the relationship between the proposed measures and subjective auditory performance assessments. In theory this should be done using carefully calibrated psycho-acoustical listening tests. We could first ask the listeners how they describe the results

with their own words, and check whether they use synonyms of “interferences” and “artifacts” or not. Then we could go on with more constrained tests, such as broadcasting pairs of results and asking listeners if they hear more or less “interferences” and “artifacts” in the first sound of each pair. Because performing these listening tests is not a trivial task, we give here only a few remarks based on our own listening experience.

If we admit that the ear splits estimated sources into the same four components than our analytical decomposition, we may define interferences, noise and artifacts as “what I hear coming from the other sources”, “what I hear coming from sensor noise” and “what I hear to be burbling artifacts”. With this definition of auditory performance measures, we remark that the SIR, SNR and SAR seem to be better related to the auditory notion of interferences, noise and artifacts using the TV Filt decomposition. Indeed decompositions using very few delays and time frames are not always able to extract all the perceived interferences inside  $e_{\text{interf}}$  but split them between  $e_{\text{interf}}$  and  $e_{\text{artif}}$ . On the contrary decompositions with too many delays and/or time frames sometimes put “burbling” artifacts into  $e_{\text{interf}}$  and nothing into  $e_{\text{artif}}$ . The parameters we chose for the TV Filt decomposition ( $F = 64$  and  $L' = 200$  ms) appear to be a good compromise in many experiments. When the TI Filt decomposition is used, a higher number of delays ( $L = 256$ ) seems preferable. Of course these choices cannot be proved using physical or mathematical arguments, but readers may check this partially by listening to the previous examples on <http://www.irisa.fr/metiss/demos/bssperf/>.

If we also admit that the ear associates energy prioritarily to the true source rather than to interferences in the case where some sources are similar, then the components  $s_{\text{target}}$ ,  $e_{\text{interf}}$ ,  $e_{\text{noise}}$  and  $e_{\text{artif}}$  estimated by the “greedy” decomposition of (9-12) should be closer to the perceptual components than those estimated by the simultaneous decomposition of (5). Indeed the “greedy” decomposition scheme takes into account similarity between sources as measured by correlation. We think that this measures part of the perceptual similarity between sources, but not all. For instance two white noises sound the same even when they are orthogonal.

Some other auditory properties cannot be explained by the proposed measures. First the high values of SIR, SNR and SAR have limited auditory signification. For example the two instantaneous mixtures of Section IV-A have very different SIRs but can hardly be distinguished. Also the SDR does not measure the total perceived distortion. In the case where  $\hat{s}_j$  is a slightly low-pass filtered version of  $s_j$ , then  $\text{SDR} \approx +\infty$  using TV Filt decomposition but low-pass filtering is perceived as timbre distortion. This fourth kind of error (besides interferences, noise and “burbling” artifacts) could be dealt with using an additional performance measure, such as Itakura-Saito distance or cepstral distortion [16].

An interesting idea to mimic the auditory treatments would be to pass the sources and noises through an auditory filter bank. Then each estimated source could be decomposed on subspaces spanned by the auditory sub-bands by handling differently sinusoidal and noise-like zones and by taking into



account auditory masking phenomena. Similar performance measures already exist in the fields of denoising and compression [13], [17].

## VI. CONCLUSION AND PERSPECTIVES

In this article we discussed the question of performance measurement in BASS. Given a set of allowed distortions, we evaluated the quality of an estimated source by four measures called SDR, SIR, SNR and SAR. Experiments involving typical mixtures and existing algorithms showed that these measures were relevant for algorithm evaluation and comparison. With respect to other existing performance measures, the main improvement is that we do not assume a particular separation algorithm nor a limited set of allowed distortions. Moreover we evaluate separately the amount of interferences, remaining sensor noise and artifacts, which is a crucial point for evaluation in under-determined mixtures.

Our performance measures are implemented within a MATLAB toolbox named *BSS\_EVAL* distributed online under the GNU Public License [18].

The main application of this work is to rank existing BASS algorithms according to their performance on the same test data. To help this we built a web database called *BASS-dB* [15] that classifies the test mixtures according to the Source Extraction subtasks [4]. These subtasks corresponds to different structures of the mixing system defined by the number of sources and sensors ( $2 \times 2$ ,  $2 \times 5$ ,  $5 \times 5$ ,  $7 \times 5$ , etc) and the kind of mixing filters (gain, delay, gain+delay, live-recorded room impulse responses). *BASS-dB* already provides some test mixtures and performance results, but we encourage people to feed it with their own mixtures and results.

We hope the BASS community will consider this issue, so that BASS methods can be compared within a shared framework. Among the results, the best BASS methods could be identified and selected for further improvement. Or objective difficulty criteria could be defined to determine for example whether the difficulty in an under-determined convolutive mixture rather comes from convolution or from under-determination. Our hypothesis that algorithm classification results are rather independent of allowed distortions could also be validated or infirmed.

Among the possible generalizations to this work, we are currently studying the derivation of psycho-acoustical performance measures and performance measures for the similar BASS tasks of Source Spatial Image Extraction and Remixing [4], that also involve listening to the extracted sources.

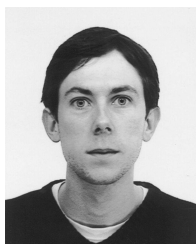
## ACKNOWLEDGMENTS

This work was part of the Junior Researchers Project "Resources for Audio Signal Separation" funded by GdR ISIS (CNRS) and was mainly performed when Emmanuel Vincent was with IRCAM, Paris (France). Cédric Févotte acknowledges support of the European Commission funded Research Training Network HASSIP (HPRN-CT-2002-00285).

The authors would like to thank Laurent Benaroya, Frédéric Bimbot, Xavier Rodet, Axel Röbel and Éric Le Carpentier for their helpful comments.

## REFERENCES

- [1] J.-F. Cardoso, "Blind source separation : statistical principles," in *Proceedings of the IEEE*, vol. 9, no. 10, oct. 1998, pp. 2009–2025, special issue on blind identification and estimation.
- [2] M. Zibulevsky and B. Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionary," *Neural Computation*, vol. 13, no. 4, 2001.
- [3] D. Ellis, "Prediction-driven computational auditory scene analysis," Ph.D. dissertation, MIT, june 1996.
- [4] E. Vincent, C. Févotte, R. Gribonval, L. Benaroya, A. Röbel, X. Rodet, F. Bimbot, and E. Le Carpentier, "A tentative typology of audio source separation tasks," in *Proc. Int. Symposium on ICA and BSS (ICA 03)*, Nara, apr. 2003, pp. 715–720.
- [5] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1-4, pp. 1–24, 2001.
- [6] D. Schobben, K. Torkkola, and P. Smaragdis, "Evaluation of blind signal separation methods," in *Proc. Int. Symposium on ICA and BSS (ICA 99)*, Aussois, jan. 1999, pp. 261–266.
- [7] R. Lambert, "Difficulty measures and figures of merit for source separation," in *Proc. Int. Symposium on ICA and BSS (ICA 99)*, Aussois, jan. 1999, pp. 133–138.
- [8] C. Févotte and C. Doncarli, "Two contributions to blind source separation using time-frequency distributions," *IEEE Signal Processing Letters*, vol. 11, no. 3, pp. 386–389, mar. 2004.
- [9] 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. Symposium on ICA and BSS (ICA 03)*, Nara, apr. 2003, pp. 993–998.
- [10] M. V. Hulle, "Clustering approach to square and non-square blind source separation," in *Proc. IEEE Workshop on Neural Networks for Signal Processing (NNSP99)*, aug. 1999, pp. 315–323.
- [11] R. Gribonval, "Sparse decomposition of stereo signals with matching pursuit and application to blind separation of more than two sources from a stereo mixture," in *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP 02)*, Orlando, may 2002.
- [12] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing  $n$  sources from 2 mixtures," in *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP 00)*, vol. 5, Istanbul, june 2000, pp. 2985–2988.
- [13] O. Cappé, "Techniques de réduction de bruit pour la restauration d'enregistrements musicaux," Ph.D. dissertation, Télécom Paris, 1993.
- [14] R. Gribonval, L. Benaroya, E. Vincent, and C. Févotte, "Proposals for performance measurement in source separation," in *Proc. Int. Symposium on ICA and BSS (ICA 03)*, Nara, apr. 2003, pp. 763–768.
- [15] R. Gribonval, E. Vincent, and C. Févotte, "BASS-dB: Blind Audio Source Separation evaluation database." [Online]. Available: <http://www.irisa.fr/metiss/BASS-dB/>
- [16] J. Deller, J. Hansen, and J. Proakis, *Discrete-time processing of speech signals*. IEEE Press, 2000.
- [17] C. Colomes, C. Schmidmer, T. Thiede, and W. Treurniet, "Perceptual quality assessment for digital audio (PEAQ) : the new ITU standard for objective measurement of perceived audio quality," in *Proc. AES 17th Int. Conf. on High Quality Audio Coding*, Firenze, sept. 1999, pp. 337–351.
- [18] C. Févotte, R. Gribonval, and E. Vincent, "BSS\_EVAL Toolbox User Guide," IRISA, Tech. Rep. 1706, 2005. [Online]. Available: [http://www.irisa.fr/metiss/bss\\_eval/](http://www.irisa.fr/metiss/bss_eval/)



**Emmanuel Vincent** graduated from École Normale Supérieure, Paris, France in 2001. He obtained the Ph.D. degree in acoustics, signal processing and computer science applied to music from the University of Paris-VI Pierre et Marie Curie, Paris, France, in 2004. He is currently a Research Assistant with the Centre for Digital Music at Queen Mary, University of London, Electronic Engineering Department, London, United Kingdom. His research focuses on structured probabilistic modeling of audio signals applied to blind source separation, indexing and

object coding of musical audio.



**Rémi Gribonval** graduated from École Normale Supérieure, Paris, France in 1997. He received the Ph.D. degree in applied mathematics from the University of Paris-IX Dauphine, Paris, France, in 1999. From 1999 until 2001 he was a visiting scholar at the Industrial Mathematics Institute (IMI) in the Department of Mathematics, University of South Carolina, SC. He is currently a Research Associate with the French National Center for Computer Science and Control (INRIA) at IRISA, Rennes, France. His research interests are in adaptive techniques for the

representation and classification of audio signals with redundant systems, with a particular emphasis in blind audio source separation.



**Cédric Févotte** was born in Laxou, France, in 1977 and lived in Tunisia, Senegal and Madagascar until 1995. He graduated from the French engineering school École Centrale de Nantes and obtained the Diplôme d'Études Approfondies en Automatique et Informatique Appliquée in 2000. He then received the Diplôme de Docteur en Automatique et Informatique Appliquée de l'École Centrale de Nantes et de l'Université de Nantes in 2003. Since november 2003 he is a Research Associate with the Signal Processing Laboratory of Cambridge University Engineering Department, United Kingdom. His current research interests concern statistical signal processing and time-frequency signal representations with application to blind source separation.