HAL
open science

The Signal Separation Evaluation Campaign (2007-2010): Achievements and Remaining Challenges

Emmanuel Vincent, Shoko Araki, Fabian J Theis, Guido Nolte, Pau Bofill, Hiroshi Sawada, Alexey Ozerov, B. Vikrham Gowreesunker, Dominik Lutter, Ngoc Q. K. Duong

To cite this version:

Emmanuel Vincent, Shoko Araki, Fabian J Theis, Guido Nolte, Pau Bofill, et al.. The Signal Separation Evaluation Campaign (2007-2010): Achievements and Remaining Challenges. [Research Report] RR-7581, 2011. inria-00579398v1

HAL Id: inria-00579398

https://inria.hal.science/inria-00579398v1

Submitted on 23 Mar 2011 (v1), last revised 11 Oct 2011 (v2)

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.

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

The Signal Separation Evaluation Campaign (2007–2010): Achievements and Remaining Challenges

Emmanuel Vincent, Shoko Araki, Fabian Theis, Guido Nolte, Pau Bofill, Hiroshi Sawada,

Alexey Ozerov, Vikrham Gowreesunker, Dominik Lutter, Ngoc Q.K. Duong

N° 7581

Mars 2011

Audio, Speech, and Language Processing

Rapport de recherche

ISSN 0249-6399 ISRN INRIA/RR--7581--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
RENNES - BRETAGNE ATLANTIQUE

The Signal Separation Evaluation Campaign (2007–2010): Achievements and Remaining Challenges

Emmanuel Vincent, Shoko Araki, Fabian Theis, Guido Nolte, Pau Bofill, Hiroshi Sawada, Alexey Ozerov, Vikrham Gowreesunker, Dominik Lutter, Ngoc Q.K. Duong

Theme : Audio, Speech, and Language Processing
Perception, Cognition, Interaction
Équipe-Projet Metiss

Rapport de recherche n° 7581 — Mars 2011 — 16 pages

Abstract: We present the outcomes of three recent evaluation campaigns in the field of audio and biomedical source separation. These campaigns have witnessed a boom in the range of applications of source separation systems in the last few years, as shown by the increasing number of datasets from 1 to 9 and the increasing number of submissions from 15 to 34. We first discuss their impact on the definition of a reference evaluation methodology, together with shared datasets and software. We then present the key results obtained over almost all datasets. We conclude by proposing directions for future research and evaluation, based in particular on the ideas raised during the related panel discussion at the Ninth International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA 2010).

Key-words: source separation, evaluation, audio, biomedical, resources

Centre de recherche INRIA Rennes – Bretagne Atlantique
IRISA, Campus universitaire de Beaulieu, 35042 Rennes Cedex
Téléphone : +33 2 99 84 71 00 — Télécopie : +33 2 99 84 71 71

La Campagne d'Évaluation de Séparation de Sources (2007–2010): Résultats et Défis Restants

Résumé : Nous présentons les résultats de trois campagnes d'évaluation récentes dans le domaine de la séparation de sources audio et biomédicales. Ces campagnes témoignent d'un élargissement du champ d'application des systèmes de séparation de sources ces dernières années, qui se traduit par une augmentation du nombre de jeux de données de 1 à 9 et du nombre de participants de 15 à 34. Nous arguons d'abord de leur impact sur la définition d'une méthodologie d'évaluation de référence, ainsi que des jeux de données et des logiciels partagés. Nous présentons ensuite les résultats clés obtenus sur presque tous les jeux de données. Nous concluons en proposant des directions pour les recherches et les évaluations à venir, basées en particulier sur les idées soulevées pendant la discussion en panel associée à la Neuvième Conférence Internationale sur l'Analyse en Variables Latentes et la Séparation de Sources (LVA/ICA 2010).

Mots-clés : séparation de sources, évaluation, audio, biomédical, ressources

1 Introduction

In many areas of signal processing, *e.g.* telecommunication, chemistry, biology and audio, the observed signals result from the combination of several sources. Source separation is the general problem of characterizing the sources and estimating the source signals underlying a given mixture signal.

Early source separation techniques based on spatial filtering are now established: beamforming and time-frequency masking are employed in mobile phones and consumer audio systems to suppress environmental noise and enhance spatial rendering [7], while independent component analysis (ICA) is used for the extraction of specific signals from electroencephalogram (EEG) and electrocardiogram (ECG) data [1]. The emergence of more powerful source separation techniques in the last five years has led to a boom in the range of applications. Data that were thought as too difficult to separate can now be processed, as illustrated by companies such as Audionamix[1] and Hit'n'Mix[2] providing commercial source separation services and software for real-world music data.

These advances have transformed source separation into a mainstream research topic, with dozens of new algorithms published every year. Regular evaluation has become necessary to reveal the effects of different algorithm designs, specify a common evaluation methodology and promote the results in other research communities and in the industry. It is with these objectives in mind that several evaluation campaigns have been held in the last few years, including the 2007 Stereo Audio Source Separation Evaluation Campaign (SASSEC) [37] and the 2008 and 2010 Signal Separation Evaluation Campaigns (SiSEC) [32, 2, 3] run by the authors in conjunction with the International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA), formerly the International Conference on Independent Component Analysis and Signal Separation.

While SASSEC was restricted to audio and fully specified by the organizers, the two SiSEC campaigns were open to all application areas and organized in a collaborative fashion. A few initial datasets, tasks and evaluation criteria were proposed by the organizers. Potential entrants were then invited to give their feedback and contribute additional specifications using collaborative software tools (wiki, mailing list). Although few people eventually took advantage of this opportunity, those who did contributed a large proportion of the evaluation materials. This resulted in the extension of the campaign to the field of biomedical signal processing, with an increasing number of datasets from 1 to 9 and an increasing number of submissions from 15 to 34. The datasets and the corresponding number of submissions for each campaign are listed in Table 1. Detailed results are available from the websites of SASSEC[3] and SiSEC[4].

In this article, we uncover the general lessons learned from these three campaigns and outline the remaining challenges. Due to the nature of the datasets, we focus on audio and to a small extent on biomedical data. The structure of the rest of the article is as follows. In Section 2, we describe the reference evaluation methodology, including shared datasets and software. In Section 3, we present the key results obtained over almost all datasets. We conclude in Section 4 by proposing directions for future research and evaluation, based in

[1] `http://www.audionamix.com/`
[2] `http://www.hitnmix.com/`
[3] `http://sassec.gforge.inria.fr/`
[4] `http://sisec.wiki.irisa.fr/`

particular on the ideas raised during the related panel discussion at LVA/ICA 2010.

Datasets	SASSEC 2007	SiSEC 2008	SiSEC 2010
Audio			
Under-determined speech and music mixtures	15	15	6
Professionally produced music recordings		9	3
Determined and over-determined mixtures		6	4
Head-mounted microphone recordings		3	2
Short two-source two-microphone recordings			7
Mixed speech and real-world background noise			6
Determined mixtures under dynamic conditions			3
Biomedical			
Cancer microarray gene expression profiles			2
EEG data with dependent components			1

Table 1: Datasets and number of submissions to the considered evaluation campaigns. The "Head-mounted microphone recordings" dataset consisted of distinct but conceptually similar recordings in 2008 and 2010, called "Head-geometry mixtures of two speech sources in real environments" and "Over-determined speech and music mixtures for human-robot interaction" respectively.

2 Reference evaluation methodology and resources

The most important outcome of SASSEC and SiSEC is perhaps the definition of a reference methodology for the evaluation of source separation systems. In particular, it has been clarified that the general problem of source separation refers to several tasks that were not always distinguished in the past. The evaluation of a source separation system requires four ingredients that we describe in the following: a dataset, a task to be addressed, one or more evaluation criteria, and ideally one or more performance bounds. Development datasets and evaluation software are available from the SiSEC website. Readers are encouraged to use these resources for the evaluation of their own systems, in order to obtain performance figures that are both reproducible and comparable with the state-of-the-art established by SiSEC.

2.1 Datasets

The datasets in Table 1 belong to two categories: *application-oriented datasets* and *diagnosis-oriented datasets*. The application-oriented datasets "Professionally produced music recordings" and "Cancer microarray gene expression profiles" consist of real-world signals, in which all the challenges underlying source separation are faced at once. The other datasets were built artificially so as to face as few challenges as possible at a time. These challenges include under-determination, *i.e.* when the number of sources is larger than the number of mixture channels, and convolutive mixing, *i.e.* when the mixing process involves

Mixture characteristics	Parameters to be specified
Sampling	Duration Sampling rate
Determination	Number of sources Number of mixture channels
Convolution (audio only)	Category of reverb *e.g.* recorded, simulated or synthetic Reverberation time Direct-to-reverberant ratio
Dynamic	Speed of the source movements Amplitude of the source movements
Sensor geometry	Relative positions of the sensors Close obstacles *e.g.* head or table (audio only)
Source geometry	Angles between the sources
Source signals	Category of sources *e.g.* male speech or adult+fetal ECG Correlation or mutual information
Noise	Category of noise *e.g.* office, cafeteria or sensor noise Input signal-to-noise ratio

Table 2: Main specifications of a diagnosis-oriented dataset.

nontrivial filters as opposed to gains or pure delays. Both categories of datasets are needed: application-oriented datasets help assessing the remaining performance gap towards industrial application, while diagnosis-oriented datasets help improving performance and robustness by combining the best solutions to individual challenges.

The characteristics of the mixtures within diagnosis-oriented datasets must be controlled as well as possible in order to quantify their difficulty. The main characteristics and the corresponding parameters to be specified are listed in Table 2. The SiSEC diagnosis-oriented audio datasets typically involve 2 to 5 different settings for each parameter of interest and as many mixture signals for each setting, so as to evaluate the effect of each setting on separation performance while fostering narrow confidence intervals on the average performance for each setting. This resulted in a total number of 27 to 84 mixtures per dataset. By contrast, the number of mixtures was limited to 5 for the application-oriented audio dataset and to a single mixture for the two biomedical datasets, for which the collection of ground truth data is notoriously harder.

2.2 Tasks and ground truth

For any data, the mixing process can always be formulated as follows [8]. Denoting by J and I the number of sources and channels, each channel $x_i(t)$, $1 \leq i \leq I$, of the mixture signal can be expressed as

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

where $s_{ij}^{\text{img}}(t)$ is the *spatial image* of source j, $1 \leq j \leq J$, on channel i, that is the contribution of this source to the observed mixture in this channel. This

Task	Ground truth	Evaluation criteria
Source counting	J	$\lvert\widehat{J} - J\rvert$
Source localization (point source)	$\theta_j(t)$	$\lvert\widehat{\theta}_j(t) - \theta_j(t)\rvert$
Mixing system estimation (point source)	$a_{ij}(t,\tau)$	MER
Source signal estimation (point source)	$s_j(t)$	SDR, SIR, SAR OPS, IPS, APS (audio)
Source spatial image estimation	$s_{ij}^{\mathrm{img}}(t)$	SDR, ISR, SIR, SAR OPS, TPS, IPS, APS (audio)
Source feature extraction	$\mathcal{F}(s_j)$ or $\mathcal{F}(s_{ij}^{\mathrm{img}})$	Depending on $\mathcal{F}$

Table 3: Main tasks, ground truth and evaluation criteria (see text for notations and acronyms).

formulation does not make any assumption on the sources, *e.g.* several distant sound sources may be considered as a single background noise source.

Under the assumption that source j is a *point source* emitting in a single spatial location, its spatial image can be further decomposed as

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

where $s_j(t)$ is a single-channel source signal and $a_{ij}(t,\tau)$ the time-varying mixing filter from source j to channel i. In the case of audio, this assumption is typically valid for speakers and small musical instruments, but not for large instruments (piano, drums) and diffuse background noise. The estimation of the mixing filters often relies on the localization of the source, expressed by its Direction-of-Arrival (DoA) $\theta_j(t)$.

Finally, in many applications, one is not interested in the source signals or the source spatial image signals themselves but in some of their *features* $\mathcal{F}(s_j)$ or $\mathcal{F}(s_{ij}^{\mathrm{img}})$. Example features include cepstral features and speech transcription in the context of noisy automatic speech recognition or the indices t of the nonzero source coefficients corresponding to active genes in the context of microarray data analysis [33, 28].

Based on the above formulation, the problem of source separation has been decomposed into six tasks listed in Table 3: *source counting*, *source spatial image estimation*[5] and *source feature extraction*, which always make sense, and *source localization*, *mixing system estimation* and *source signal estimation*, which make sense for point sources only. Each task corresponds to a distinct quantity to be estimated.

Evaluation consists of comparing the estimated quantity with the *ground truth* according to one or more criteria. The way to obtain the ground truth data depends whether the dataset consists of synthetic or recorded mixtures.

[5]The task of estimating the subspace spanned by certain point sources, which was specified for the EEG dataset in SiSEC 2010, is formally equivalent to the estimation of the spatial image of these sources considered as a single diffuse source.

Ground truth data are typically available for all tasks in the former case but not in the latter. One popular technique for the acquisition of ground truth data for *live audio recordings* consists of separately recording each source in turn, thus yielding ground truth source spatial image signals, and summing them to obtain the mixture signal [29]. This approach cannot be used for real-world biomedical datasets, for which the sources cannot be switched off. When feasible, the ground truth is then specified by experts.

2.3 Evaluation criteria

The evaluation criteria for source counting and source localization are straightforward. The evaluation of mixing system estimation is less obvious. Evaluation criteria for over-determined mixing systems [21] and a Mixing Error Ratio (MER) criterion [32] can be found in the literature, but their correlation with source separation performance is unclear, so that agreed-upon criteria remain to be defined.

A few evaluation criteria have been proposed for source signal estimation and source spatial image estimation. Early criteria [21, 39] were restricted to linear unmixing or binary time-frequency masking and required knowledge of the unmixing filters or the time-frequency masks. More recently, a family of criteria has been proposed that applies to all audio mixtures and algorithms [34, 37]. In the case of source spatial image estimation, the criteria derive from the decomposition of an estimated source image $\hat{s}_{ij}^{\text{img}}(t)$ as [37]

$$\hat{s}_{ij}^{\text{img}}(t) = s_{ij}^{\text{img}}(t) + e_{ij}^{\text{spat}}(t) + e_{ij}^{\text{interf}}(t) + e_{ij}^{\text{artif}}(t) \tag{3}$$

where $s_{ij}^{\text{img}}(t)$ is the true source image and $e_{ij}^{\text{spat}}(t)$, $e_{ij}^{\text{interf}}(t)$ and $e_{ij}^{\text{artif}}(t)$ are distinct error components representing spatial (or filtering) distortion, interference and artifacts. This decomposition is motivated by the perceptual distinction between sounds from the target source, sounds from other sources and "musical noise", corresponding to the signals $s_{ij}^{\text{img}}(t) + e_{ij}^{\text{spat}}(t)$, $e_{ij}^{\text{interf}}(t)$ and $e_{ij}^{\text{artif}}(t)$ respectively. Spatial distortion and interference components are expressed as filtered versions of the true source images, computed by least-squares projection of the estimated source image onto the corresponding signal subspaces

$$e_{ij}^{\text{spat}}(t) = P_j^L(\hat{s}_{ij}^{\text{img}})(t) - s_{ij}^{\text{img}}(t) \tag{4}$$

$$e_{ij}^{\text{interf}}(t) = P_{\text{all}}^L(\hat{s}_{ij}^{\text{img}})(t) - P_j^L(\hat{s}_{ij}^{\text{img}})(t) \tag{5}$$

$$e_{ij}^{\text{artif}}(t) = \hat{s}_{ij}^{\text{img}}(t) - P_{\text{all}}^L(\hat{s}_{ij}^{\text{img}})(t) \tag{6}$$

where P_j^L is the least-squares projector onto the subspace spanned by $s_{kj}^{\text{img}}(t-\tau)$, $1 \leq k \leq I$, $0 \leq \tau \leq L-1$, P_{all}^L is the least-squares projector onto the subspace spanned by $s_{kl}^{\text{img}}(t-\tau)$, $1 \leq k \leq I$, $1 \leq l \leq J$, $0 \leq \tau \leq L-1$, and the filter length L is set to 32 ms. The amount of spatial distortion, interference and artifacts is then measured by three energy ratios expressed in decibels (dB): the *source Image to Spatial distortion Ratio* (ISR), the *Signal to Interference Ratio*

(SIR) and the *Signal to Artifacts Ratio* (SAR)

$$\mathrm{ISR}_j = 10\log_{10}\frac{\sum_{i=1}^{I}\sum_t s_{ij}^{\mathrm{img}}(t)^2}{\sum_{i=1}^{I}\sum_t e_{ij}^{\mathrm{spat}}(t)^2} \tag{7}$$

$$\mathrm{SIR}_j = 10\log_{10}\frac{\sum_{i=1}^{I}\sum_t (s_{ij}^{\mathrm{img}}(t)+e_{ij}^{\mathrm{spat}}(t))^2}{\sum_{i=1}^{I}\sum_t e_{ij}^{\mathrm{interf}}(t)^2} \tag{8}$$

$$\mathrm{SAR}_j = 10\log_{10}\frac{\sum_{i=1}^{I}\sum_t (s_{ij}^{\mathrm{img}}(t)+e_{ij}^{\mathrm{spat}}(t)+e_{ij}^{\mathrm{interf}}(t))^2}{\sum_{i=1}^{I}\sum_t e_{ij}^{\mathrm{artif}}(t)^2}. \tag{9}$$

The total error is also measured by the *Signal to Distortion Ratio* (SDR)

$$\mathrm{SDR}_j = 10\log_{10}\frac{\sum_{i=1}^{I}\sum_t s_{ij}^{\mathrm{img}}(t)^2}{\sum_{i=1}^{I}\sum_t (e_{ij}^{\mathrm{spat}}(t)+e_{ij}^{\mathrm{interf}}(t)+e_{ij}^{\mathrm{artif}}(t))^2} \tag{10}$$

In the case of source signal estimation, similar criteria can be defined by grouping the first two terms in (3) [34]. Indeed, the source signals can only be estimated up to arbitrary filtering, which should not be taken into account in the SDR. Similarly, when the sources are estimated in arbitrary order, the order is selected that leads to the largest average SIR. Improved auditory-motivated variants of these criteria termed Target-related Perceptual Score (TPS), Interference-related Perceptual Score (IPS), Artifact-related Perceptual Score (APS) and Overall Perceptual Score (OPS) have also been employed [16].

Finally, the evaluation criteria related to source feature extraction depend on the considered features. Noisy automatic speech recognition may be evaluated in terms of Word Error Rate (WER) while the detection of the indices t of the nonzero source coefficients in the context of microarray data analysis may be evaluated by counting the number of significantly detected indices using appropriate statistical tests [3].

2.4 Baseline algorithms and performance bounds

In addition to quantifying the performance of the source separation system under test, it is recommended to evaluate some reference algorithms via the same criteria. Indeed, the performance of all systems varies a lot depending on the mixture signal, so that the difference of performance with respect to reference algorithms often provides a more robust indicator. Two categories of reference algorithms have been considered in SiSEC: baseline algorithms providing medium to poor performance and *oracle estimators* providing theoretical upper bounds on performance. A range of oracle estimators were defined in [35, 38] for linear unmixing-based and time-frequency masking-based algorithms.

3 Key results

3.1 Audio source separation

The audio datasets of SASSEC and SiSEC attracted a total of 79 submissions, from which many useful conclusions can be drawn. We let readers refer to [37, 32, 2] for the detailed performance of each system as a function of the

mixture characteristics, and provide here a broader perspective over the field by focusing on the best systems on average. Furthermore, we concentrate on the source signal estimation and source spatial image estimation tasks, which are the only ones for which sufficient submissions are available.

3.1.1 Evolution of performance over the "Under-determined speech and music mixtures" dataset

We first analyze the evolution of performance over the only dataset that was considered within the three campaigns, that is the "Under-determined speech and music mixtures" dataset. Due to the evolution of the dataset itself, the source signals were different in SASSEC and SiSEC. In order to compare the results, we consider the same categories of mixtures in both cases, that is two 2-channel mixtures of 4 speech sources and two 2-channel mixtures of 3 music sources mixed in three different ways: instantaneous mixing, live recording with 250 ms reverberation time and 5 cm microphone spacing, and live recording with 250 ms reverberation time and 1 m microphone spacing. For each campaign and each mixing condition, we select the system leading to best average SDR over all sources and all mixtures[6].

The resulting average SDR, ISR, SIR and SAR are reported in Table 4. The separation of instantaneous mixtures is close to be solved in 2010, with an average SDR of 14 dB, while that of live recordings remains much more difficult, with an average SDR of 3 dB. All performance criteria improved by 3 to 4 dB on instantaneous mixtures when replacing the Sparse Component Analysis (SCA) method in [31] by multichannel Nonnegative Matrix Factorization (NMF) [24] or by the flexible probabilistic modeling framework in [25]. These new methods are examples of the emerging variance modeling framework [36] for audio source separation, which addresses some shortcomings of the conventional linear modeling framework [19] underlying ICA and SCA by enabling the exploitation of additional prior information about the source spectra. These methods remain inferior to conventional SCA on live recordings, however, perhaps because of the omnipresence of local optima in the objective function and the need for more accurate initialization. The best current SDR on these live recordings [27] remains 6 dB below that of the binary masking oracle [35], which indicates that room is left for progress.

3.1.2 Current performance on the other audio datasets

In addition to the above dataset which was used for all campaigns, two datasets, namely "Professionally produced music recordings" and "Determined and over-determined mixtures", were used for the last two campaigns. The corresponding results do not reveal any performance increase, however, but a performance decrease instead, due to the fact that different methods were submitted in 2008 and 2010.

The current performance on these two datasets and on the remaining audio datasets is shown in Table 5. For each dataset, we select the method providing

[6]The choice of the best system depends on the eventual application scenario, since different applications may involve different mixture characteristics and different evaluation criteria. Our choice promotes versatile algorithms that were able to separate all sources within all mixtures of the dataset.

Performance	SASSEC 2007	SiSEC 2008	SiSEC 2010	Binary masking oracle 2008 and 2010
Instantaneous mixtures				
Method	[31]	[24]	[25]	[35]
SDR (dB)	10.3	14.0	13.4	10.4
ISR (dB)	19.2	23.3	23.4	19.4
SIR (dB)	16.0	20.4	20.0	21.1
SAR (dB)	12.2	15.4	14.9	11.4
Live recordings with 5 cm microphone spacing				
Method	[27]	[14]	[27]	[35]
SDR (dB)	1.8	2.6	3.5	9.2
ISR (dB)	7.0	5.7	8.4	16.9
SIR (dB)	4.2	2.4	7.0	18.5
SAR (dB)	6.8	7.3	6.3	9.9
Live recordings with 1 m microphone spacing				
Method	[27]	[14]	[27]	[35]
SDR (dB)	3.6	2.5	3.2	9.1
ISR (dB)	8.4	5.8	8.1	16.6
SIR (dB)	6.9	2.9	6.6	18.2
SAR (dB)	6.8	7.3	6.4	9.8

Table 4: Evolution of the average performance of the best source spatial image estimation method over the "Under-determined speech and music mixtures" dataset compared to that of the binary masking oracle.

the best average SDR over all sources of all mixtures, except for the "Determined and over-determined mixtures" dataset for which we consider the SIR instead[7], and for the "Head-mounted microphone recordings" datasets for which the best method separated only two sources out of three.

The best separation is achieved on noiseless over-determined mixtures, with an average SIR of 14 dB for 5-channel recordings of 3 sources and 11 dB for 4-channel recordings of 2 sources. The corresponding methods both rely on frequency-domain ICA, where the source signals estimated within each frequency bin are ordered based either on their spatial location [22] or on the correlation of their temporal activity patterns [26]. Similar performance is achieved over 2-channel noiseless mixtures of 2 sources, again by means of frequency-domain ICA [23]. Note that the considered 2-channel 2-source mixtures were either short or dynamic, which shows that frequency-domain ICA can efficiently adapt to such situations [23]. These methods result in significant residual convolution of the source signals, however, as indicated by the lower SAR.

Performance drops on 4-channel mixtures of 4 sources, for which the best 4-channel 2-source separation method [22] achieves a SIR of 3 dB only, and on professionally produced music recordings, for which the best method [4] based on the aforementioned variance modeling framework provided a SIR of 9 dB.

[7]Due to the unavailability of the ground truth source signals in this dataset, the results of the source signal estimation task were evaluated with respect to the first channel of the spatial image of each source instead. Only the SIR criterion then makes sense according to the specification of the task in Section 2.2.

Dataset	Number of channels and sources	Method	SDR (dB)	ISR (dB)	SIR (dB)	SAR (dB)
SiSEC 2008						
Professionally produced music recordings	$I = 2$ $J = 2$ to 10	[4]	4.9	9.9	8.6	7.8
Determined and over-determined mixtures	$I = 4$ $J = 2$	[22]	N/A	N/A	11.9	N/A
	$I = 4$ $J = 4$				3.1	
SiSEC 2010						
Head-mounted microphone recordings	$I = 5$ $J = 3$	[26]	1.7	N/A	14.3	2.5
Short two-source two-microphone recordings	$I = 2$ $J = 2$	[23]	5.9	10.3	11.4	17.1
Mixed speech and real-world background noise	$I = 2$ $J = 1$	[13]	2.7	16.1	4.4	11.9
	$I = 4$ $J = 1$	[26]	7.5	17.1	10.4	14.3
Determined mixtures under dynamic conditions	$I = 2$ $J = 2$	[23]	6.2	N/A	13.8	7.4

Table 5: Average performance of the best source separation method over all audio datasets except the "Under-determined speech and music mixtures" dataset. Figures relate to the source spatial image estimation task when the ISR is reported and to the source signal estimation task otherwise.

This suggests that performance does not depend so much whether the mixture is determined or over-determined but rather on the number of sources itself, since a larger number of sources makes it more difficult to achieve accurate source localization, which is a prerequisite in most source separation methods. The presence of background noise appears even more detrimental. Indeed, the SIR decreases by 8 dB when replacing one of the sources within a 2-channel 2-source mixture by diffuse background noise, yielding a SIR as low as 4 dB. This appears due to the lack of accurate noise models, despite recent advances in this direction in [13].

Finally, it must be emphasized that none of the above methods is truly blind. All methods assume prior knowledge of the number of sources and the category of mixing (instantaneous vs. convolutive), and most submissions to the "Professionally produced music recordings" dataset even relied on manual parameter fixing or manual grouping of the sounds composing each source.

3.2 Biomedical source separation

Fewer conclusions can be drawn from the biomedical source separation results in SiSEC 2010, due to the smaller number of submissions. We summarize here the results obtained over the microarray gene expression dataset.

In this context, each channel $x_i(t)$ of the mixture signal, called expression profile, measures the level of messenger RNA (mRNA) corresponding to one gene t within one subject or experimental condition i. The expression profiles can be regarded as a linear instantaneous mixture of several cell signaling

pathways or more generally biological processes [18, 17]. Using source separation techniques, the estimated source signals can be interpreted as patterns reflecting active signaling pathways. In SiSEC 2010, mRNA was extracted from $I = 189$ invasive breast carcinomas, measured using Affymetrix U133A genechips and normalized via the RMA algorithm. Non-expressed genes were filtered out, resulting in a total of $T = 11815$ expressed genes [3]. The $J = 10$ ground truth signaling pathways were approximated as simple gene lists, taken from NETPATH[8]. We evaluated the quality of the estimated pathways by means of statistical tests [3]. More precisely, for each source signal, we identified the genes mapping to the distinct pathways and calculated p-values using Fisher's exact test. We then used the Benjamini-Hochberg procedure to correct for multiple testing and declared an estimated pathway as enriched if its p-value was below 0.05. We finally counted the total number of distinct enriched pathways.

Two methods were submitted that both rely on some form of prior information, implemented either via matrix factorization using a graph model (GraDe) [6] or via Network Component Analysis (NCA) [9]. For each of the 10 ground truth pathways, both methods found at least one matching pathway with a p-value below 0.05 according to Fisher's exact test. After discarding duplicate pathways, the number of correctly estimated pathways reduced to 7 and 5, respectively. Finally, after Benjamini-Hochberg correction, the number of enriched pathways was equal to 5 and 0, respectively. This shows that the GraDe approach clearly outperformed the NCA approach. We hypothesize that the better performance of GraDe arises from the inclusion of pathway information within the graph model.

4 Remaining challenges

To sum up, SASSEC and SiSEC have been instrumental in the definition of a clear evaluation methodology for audio and biomedical source separation and in the creation of data and software resources. The results support the emergence of source separation systems exploiting advanced source models accounting for the source spectra in the case of audio source separation [24, 25, 36, 4] or for signaling pathway information in the case of biomedical source separation [6]. Nevertheless, more conventional methods based on frequency-domain ICA or SCA still perform best on live audio recordings of many sources and/or background noise [27, 22, 26, 23].

4.1 Evaluation methodology

The biggest challenge regarding evaluation methodology consists of extending the methodology summarized in this article to other datasets, tasks and application domains. Up to 2010, SASSEC and SiSEC have mainly focused on audio source signal estimation and source spatial image estimation, which are perhaps not the most useful tasks in the real world, and left the other audio tasks [33] aside. Recently, a comprehensive dataset has been created for the evaluation of audio source separation systems in terms of WER in the context of noise-robust speech recognition in a domestic environment [12, 10]. Stereo to multichannel upmix [5] is also a vibrant area of research to which advanced source separation

[8] `http://www.netpath.org`

systems could contribute and for which novel performance criteria are needed. Appropriate statistical confidence measures, tighter oracle performance bounds and advanced diagnosis procedures such as those in [11, 15, 20] are also needed to increase the insight that can be gained from evaluation. Finally, increased efforts should be made to promote source separation evaluations in the biomedical signal processing community, as well as in other communities, *e.g.* cosmology or telecommunications, where the proposed tasks and evaluation criteria might also apply.

4.2 Key challenges for future research

In addition to these methodological challenges, we identified three key challenges for future research in audio and biomedical source separation in light of the campaign results:

- the experimentation of advanced source models and mixing models including as much available information as possible, especially for complex sources such as nonstationary background noise or taking into account the wealth of prior information readily available in the biomedical context,
- the design of accurate source localization methods, which are required for parameter initialization of the mixing model, especially for short and/or dynamic mixtures,
- the development of model selection techniques enabling truly blind separation by automatically finding the number of sources and adapting the source models and the mixing model to the mixture at hand.

Although recent advances have been made in each of these directions [25, 6, 22, 30], they remain to be fully developed, combined together and validated on real-world data.

5 Acknowledgments

We would like to thank all the entrants and all the persons besides the authors who helped organizing SiSEC 2008 and 2010 by contributing datasets, code or part of their time (in alphabetical order): J. Anemüller, M. Durkovic, M. Dyrholm, V. Emiya, K. E. Hild II, N. Ito, H. Kayser, M. Kleinsteuber, Z. Koldovsky, O. Le Blouch, B. Lösch, F. Nesta, L. C. Parra, M. Rothbucher, H. Shen, P. Tichavsky, M. Vinyes Raso and J. Woodruff.

References

[1] L. Albera, A. Kachenoura, A. Karfoul, P. Comon, and L. Senhadji. One decade of biomedical problems using ICA: a full comparative study. In *Proc. 2009 World Congress on Medical Physics and Biomedical Engineering*, pages 2269–2272, 2009.

[2] S. Araki, A. Ozerov, V. Gowreesunker, H. Sawada, F. Theis, G. Nolte, D. Lutter, and N. Q. K. Duong. The 2010 Signal Separation Evaluation

Campaign (SiSEC 2010): Audio source separation. In *Proc. 9th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pages 114–122, 2010.

[3] S. Araki, F. Theis, G. Nolte, D. Lutter, A. Ozerov, V. Gowreesunker, H. Sawada, and N. Q. K. Duong. The 2010 Signal Separation Evaluation Campaign (SiSEC 2010): Biomedical source separation. In *Proc. 9th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pages 123–130, 2010.

[4] S. Arberet, A. Ozerov, F. Bimbot, and R. Gribonval. A tractable framework for estimating and combining spectral source models for audio source separation. *Signal Processing*, page (submitted), 2011.

[5] C. Avendano and J.-M. Jot. A frequency-domain approach to multichannel upmix. *Journal of the Audio Engineering Society*, 52(7/8):740–749, 2004.

[6] F. Blöchl, A. Kowarsch, and F. J. Theis. Second-order source separation based on prior knowledge realized in a graph model. In *Proc. 9th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pages 434–441, 2010.

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

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

[9] W. Chen, C. Q. Chang, and Y. S. Hung. Transcription factor activity estimation based on particle swarm optimization and fast network component analysis. In *Proc. Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1061–1064, 2010.

[10] H. Christensen, J. Barker, N. Ma, and P. Green. The CHiME corpus: a resource and a challenge for computational hearing in multisource environments. In *Proc. Interspeech*, pages 1918–1921, 2010.

[11] W. J. Conover. *Practical Non-Parametric Statistics*. Wiley, 1980.

[12] M. Cooke, J. R. Hershey, and S. J. Rennie. Monaural speech separation and recognition challenge. *Computer Speech and Language*, 24(1):1–15, 2010.

[13] N. Q. K. Duong, E. Vincent, and R. Gribonval. Under-determined reverberant audio source separation using local observed covariance and auditory-motivated time-frequency representation. In *Proc. 9th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pages 73–80, 2010.

[14] Z. El Chami, A. D.-T. Pham, C. Servière, and A. Guerin. A new model based underdetermined source separation. In *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2008.

[15] V. Emiya, E. Vincent, and R. Gribonval. An investigation of discrete-state discriminant approaches to single-sensor source separation. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 97–100, 2009.

[16] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann. Subjective and objective quality assessment of audio source separation. *IEEE Transactions on Audio, Speech and Language Processing*, 2011. (to appear).

[17] S.-I. Lee and S. Batzoglou. Application of independent component analysis to microarrays. *Genome Biology*, 4:R76, 2003.

[18] W. Liebermeister. Linear modes of gene expression determined by independent component analysis. *Bioinformatics*, 18(1):51–60, 2002.

[19] S. Makino, T.-W. Lee, and H. Sawada, editors. *Blind speech separation*. Springer, 2007.

[20] M. I. Mandel, S. Bressler, B. Shinn-Cunningham, and D. P. W. Ellis. Evaluating source separation algorithms with reverberant speech. *IEEE Transactions on Audio, Speech and Language Processing*, 18(7):1872–1883, 2010.

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

[22] F. Nesta, M. Omologo, and P. Svaizer. Multiple TDOA estimation by using a state coherence transform for solving the permutation problem in frequency-domain BSS. In *Proc. IEEE Workshop on Machine Learning for Signal Processing (MLSP)*, pages 43–48, 2008.

[23] F. Nesta, P. Svaizer, and M. Omologo. Convolutive BSS of short mixtures by ICA recursively regularized across frequencies. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(3):624–639, 2011.

[24] A. Ozerov and C. Févotte. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):550–563, 2010.

[25] A. Ozerov, E. Vincent, and F. Bimbot. A general flexible framework for the handling of prior information in audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 2010. (submitted).

[26] H. Sawada, S. Araki, and S. Makino. Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS. In *Proc. IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 3247–3250, 2007.

[27] H. Sawada, S. Araki, and S. Makino. Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(3):516–527, 2011.

[28] R. Schachtner, D. Lutter, P. Knollmüller, A. M. Tomé, F. J. Theis, G. Schmitz, M. Stetter, P. Gómez Vilda, and E. W. Lang. Knowledge-based gene expression classification via matrix factorization. *Bioinformatics*, 24:1688–1697, 2008.

[29] D. Schobben, K. Torkkola, and P. Smaragdis. Evaluation of blind signal separation methods. In *Proc. 1st Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA)*, pages 261–266, 1999.

[30] V. Y. F. Tan and C. Févotte. Automatic relevance determination in non-negative matrix factorization. In *Proc. Workshop on Signal Processing with Adaptative Sparse Structured Representations (SPARS)*, 2009.

[31] E. Vincent. Complex nonconvex l_p norm minimization for underdetermined source separation. In *Proc. 7th Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, pages 430–437, 2007.

[32] E. Vincent, S. Araki, and P. Bofill. The 2008 Signal Separation Evaluation Campaign: A community-based approach to large-scale evaluation. In *Proc. 8th Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, pages 734–741, 2009.

[33] E. Vincent, C. Févotte, R. Gribonval, L. Benaroya, X. Rodet, A. Röbel, É. Le Carpentier, and F. Bimbot. A tentative typology of audio source separation tasks. In *Proc. 4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA)*, pages 715–720, 2003.

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

[35] E. Vincent, R. Gribonval, and M. D. Plumbley. Oracle estimators for the benchmarking of source separation algorithms. *Signal Processing*, 87(8):1933–1950, 2007.

[36] E. Vincent, M. G. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies. Probabilistic modeling paradigms for audio source separation. In *Machine Audition: Principles, Algorithms and Systems*, pages 162–185. IGI Global, 2010.

[37] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca. First Stereo Audio Source Separation Evaluation Campaign: Data, algorithms and results. In *Proc. 7th Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, pages 552–559, 2007.

[38] D. L. Wang. On ideal binary mask as the computational goal of auditory scene analysis. In *Speech Separation by Humans and Machines*, pages 181–197. Springer, New York, NY, 2005.

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

Centre de recherche INRIA Rennes – Bretagne Atlantique
IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Grenoble – Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Saclay – Île-de-France : Parc Orsay Université - ZAC des Vignes : 4, rue Jacques Monod - 91893 Orsay Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399