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 Duong

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

Emmanuel Vincent[a,*], Shoko Araki[b], Fabian Theis[c], Guido Nolte[d], Pau Bofill[e], Hiroshi Sawada[b], Alexey Ozerov[a,1], Vikrham Gowreesunker[f], Dominik Lutter[c], Ngoc Q.K. Duong[a]

[a] *INRIA, Centre de Rennes - Bretagne Atlantique, 35042 Rennes Cedex, France*
[b] *NTT Communication Science Labs, NTT Corporation, 2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan*
[c] *Institute for Bioinformatics and Systems Biology, Helmholtz Zentrum München, Ingolstädter Landstraße, 85764 Neuherberg, Germany*
[d] *Fraunhofer FIRST.IDA, Kekuléstrasse 7, 12489 Berlin, Germany*
[e] *Department of Computer Architecture, Universitat Politècnica de Catalunya, Campus Nord Mòdul D6, Jordi Girona 1-3, 08034 Barcelona, Spain*
[f] *DSP Solutions R&D Center, Texas Instruments Inc., 12500 TI Boulevard, MS 8649, Dallas, TX 75243, USA*

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).

Keywords: source separation, evaluation, audio, biomedical, resources

*Corresponding author. Tel.: +33 2 9984 2269; Fax.: +33 2 9984 7171
Email address: emmanuel.vincent@inria.fr (Emmanuel Vincent)

[1] Alexey Ozerov was supported by the Quaero Programme, funded by OSEO.

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 [1, 2], while independent component analysis (ICA) is used for the extraction of specific signals from electroencephalogram (EEG), electrocardiogram (ECG) and functional magnetic resonance imaging (fMRI) data [3, 4]. 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 providing commercial source separation services and software for real-world music data [5].

These advances have transformed source separation into a mainstream research topic, with dozens of new algorithms published every year[2]. Regular evaluation has become necessary to reveal the effects of different algorithm designs, specify a common evaluation methodology and promote new 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) [6] and the 2008 and 2010 Signal Separation Evaluation Campaigns (SiSEC) [7, 8, 9] run by the authors in conjunction with the 2007 and 2009 International Conferences on Independent Component Analysis and Signal Separation (ICA) and the 2010 International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA).

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

[2] According to the Google Scholar search engine, the annual number of publications containing the words "audio" and "source separation" has roughly doubled every 2 years from 14 in 1994 to 903 in 2007 and has stalled around 900 since then. The annual number of publications containing the words "microarray" and "source separation" has roughly doubled every 4 years from 14 in 2002 to 76 in 2010 and is still increasing.

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 an increasing number of datasets from 1 to 9 and an increasing number of submissions from 15 to 34 and in the extension of the campaign to the field of biomedical signal processing. 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 particular on the ideas raised during the related panel discussion at LVA/ICA 2010.

Datasets	Number of submissions		
	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: Number of submissions associated to each dataset for each of 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.

[3] `http://sassec.gforge.inria.fr/`

[4] `http://sisec.wiki.irisa.fr/`

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,
- ideally, one or more performance bounds.

Development datasets and evaluation software are available from the SiSEC website[4]. 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,
- 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 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 applications, 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.

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

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

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 multichannel test signals for each setting, so as to evaluate the effect of each setting on separation performance while favoring narrow confidence intervals on the average performance for each setting. This resulted in a total number of 27 to 84 test signals per dataset. By contrast, the number of test signals was limited to 5 for the application-oriented audio dataset and to a single test signal 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 [10]. 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 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

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

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 $\mathcal{F}(s_{ij}^{\text{img}})$. 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 [11, 12].

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. 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 [13]. This approach cannot be used for

[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 [10].

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 ISI, PI (over-determined)
Source signal estimation (point source)	$s_j(t)$	SDR, SIR, SAR OPS, IPS, APS (audio)
Source spatial image estimation	$s_{ij}^{\text{img}}(t)$	SDR, ISR, SIR, SAR OPS, TPS, IPS, APS (audio)
Source feature extraction	$\mathcal{F}(s_j)$ or $\mathcal{F}(s_{ij}^{\text{img}})$	Depending on $\mathcal{F}$

Table 3: Main tasks, ground truth and evaluation criteria. See Section 2.2 for the notations and Section 2.3 for the acronyms.

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.

Regarding the evaluation of mixing system estimation, several established criteria such as the *Amari Performance Index* (PI) or the *Inter-Symbol Interference* (ISI) have been proposed for over-determined mixing systems [14, 15] and widely applied to biomedical data. A more general *Mixing Error Ratio* (MER) criterion applicable to all mixing systems has been introduced in [7]. For instantaneous mixtures, the estimated mixing gains $\widehat{a}_{ij}$ for a given source j are decomposed as

$$\widehat{a}_{ij} = a_{ij}^{\text{coll}} + a_{ij}^{\text{orth}} \tag{3}$$

where a_{ij}^{coll} and a_{ij}^{orth} are respectively collinear and orthogonal to the true vector of mixing gains a_{ij}, $1 \leq i \leq I$, and are computed by least squares projection. Accuracy is then assessed via the following ratio in decibels (dB)

$$\text{MER}_j = 10 \log_{10} \frac{\sum_{i=1}^{I} a_{ij}^{\text{coll}\,2}}{\sum_{i=1}^{I} a_{ij}^{\text{orth}\,2}}. \tag{4}$$

More generally, for time-varying convolutive mixtures, the accuracy of estimated mixing filters for source j is assessed by computing the MER in

each frequency bin ν between $\widehat{a}_{ij}(t,\nu)$ and $a_{ij}(t,\nu)$ and averaging it over frequency and time. When the sources are estimated in arbitrary order, the order is selected that leads to the largest average MER. This criterion has been little used so far, however, so that general agreed-upon criteria remain to be found.

Several evaluation criteria have also been proposed for source signal estimation and source spatial image estimation. Early criteria applied to biomedical data or toy audio data [13, 14, 16] 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 mixtures and algorithms [17, 6]. 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 [6]

$$\hat{s}_{ij}^{\text{img}}(t) = s_{ij}^{\text{img}}(t) + e_{ij}^{\text{spat}}(t) + e_{ij}^{\text{interf}}(t) + e_{ij}^{\text{artif}}(t) \tag{5}$$

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 distinction between signal from the target source, residual noise from the other sources and extraneous noise introduced by the algorithm[6], 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

$$e_{ij}^{\text{spat}}(t) = P_j^L(\hat{s}_{ij}^{\text{img}})(t) - s_{ij}^{\text{img}}(t) \tag{6}$$

$$e_{ij}^{\text{interf}}(t) = P_{\text{all}}^L(\hat{s}_{ij}^{\text{img}})(t) - P_j^L(\hat{s}_{ij}^{\text{img}})(t) \tag{7}$$

$$e_{ij}^{\text{artif}}(t) = \hat{s}_{ij}^{\text{img}}(t) - P_{\text{all}}^L(\hat{s}_{ij}^{\text{img}})(t) \tag{8}$$

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 $s_{kl}^{\text{img}}(t-\tau)$, $1 \leq k \leq I$, $1 \leq l \leq J$, $0 \leq \tau \leq L-1$. The length L of the distortion filter is equal to 1 tap for EEG, ECG or fMRI and is typically set to 32 ms in an audio context. 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*

[6] In the context of audio, this extraneous noise is called "musical noise".

(ISR), the *Signal to Interference Ratio* (SIR) and the *Signal to Artifacts Ratio* (SAR)

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

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

$$\mathrm{SAR}_j = 10\log_{10}\frac{\sum_{i=1}^{I}\sum_t (s_{ij}^{\mathrm{img}}(t)+e_{ij}^{\mathrm{spat}}(t)+e_{ij}^{\mathrm{interf}}(t))^2}{\sum_{i=1}^{I}\sum_t e_{ij}^{\mathrm{artif}}(t)^2}. \tag{11}$$

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

$$\mathrm{SDR}_j = 10\log_{10}\frac{\sum_{i=1}^{I}\sum_t s_{ij}^{\mathrm{img}}(t)^2}{\sum_{i=1}^{I}\sum_t (e_{ij}^{\mathrm{spat}}(t)+e_{ij}^{\mathrm{interf}}(t)+e_{ij}^{\mathrm{artif}}(t))^2} \tag{12}$$

In the case of source signal estimation, similar criteria can be defined by grouping the first two terms in (5) [17]. 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. In the specific case of audio, 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 [18].

Finally, the evaluation criteria related to source feature extraction are highly specific to the considered features. For example, 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, as detailed in Section 3.2.

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 [19, 20] 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 [6, 7, 8] 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[7].

The resulting average SDR, ISR, SIR and SAR are reported in Table 4. The following observations can be made:

- 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.

[7]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	[21]	[22]	[23]	[19]
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	[25]	[26]	[25]	[19]
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	[25]	[26]	[25]	[19]
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.

- All performance criteria improved by 3 to 4 dB on instantaneous mixtures when replacing the Sparse Component Analysis (SCA) method in [21] by multichannel Nonnegative Matrix Factorization (NMF) [22] or by the flexible probabilistic modeling framework in [23]. These new methods are examples of the emerging *variance modeling* framework [24] for audio source separation. This framework addresses some shortcomings of the conventional linear modeling framework [2] underlying ICA and SCA by enabling the exploitation of additional prior information about the source spectra.

- These new 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 [25] remains 6 dB below that of the binary masking oracle [19], 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 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[8], and for the "Head-mounted microphone recordings" datasets for which the best method separated only two sources out of three. The following observations can be made:

- Not surprisingly, 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 [27] or on the correlation of their temporal activity patterns [28].

- Similar performance is achieved over 2-channel noiseless mixtures of 2 sources, again by means of frequency-domain ICA [29]. 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 [29]. These methods result in significant filtering distortion 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 [27] achieves a SIR of 3 dB only, and on professionally produced music recordings, for which the best method [30] based on the aforementioned variance modeling

[8] 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	[30]	4.9	9.9	8.6	7.8
Determined and over-determined mixtures	$I = 4$ $J = 2$	[27]	N/A	N/A	11.9	N/A
	$I = 4$ $J = 4$				3.1	
SiSEC 2010						
Head-mounted microphone recordings	$I = 5$ $J = 3$	[28]	1.7	N/A	14.3	2.5
Short two-source two-microphone recordings	$I = 2$ $J = 2$	[29]	5.9	10.3	11.4	17.1
Mixed speech and real-world background noise	$I = 2$ $J = 1$	[31]	2.7	16.1	4.4	11.9
	$I = 4$ $J = 1$	[28]	7.5	17.1	10.4	14.3
Determined mixtures under dynamic conditions	$I = 2$ $J = 2$	[29]	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.

framework provided a SIR of 9 dB. 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 [31].

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 ribonucleic acid (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 [32, 33]. 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 gene-chips and normalized via the robust multi-array average (RMA) algorithm. Non-expressed genes were filtered out, resulting in a total of $T = 11815$ expressed genes [9]. The $J = 10$ ground truth signaling pathways were approximated as simple gene lists, taken from NETPATH[9]. The quality of the estimated pathways was evaluated by means of statistical tests [9]. More precisely, for each source signal, the genes mapping to the distinct pathways were identified and p-values were calculated using Fisher's exact test. The Benjamini-Hochberg procedure was then used to correct for multiple testing and an estimated pathway was declared as enriched if its p-value was below 0.05. Finally, the total number of distinct enriched pathways was counted.

Two methods were submitted that both rely on some form of prior information, implemented either via matrix factorization using a graph model (GraDe) [34] or via Network Component Analysis (NCA) [35]. 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.

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

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 [22, 23, 24, 30] or for signaling pathway information in the case of biomedical source separation [34]. 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 [25, 27, 28, 29].

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 [11] 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 [36, 37]. Stereo to multichannel upmix [38] is also a vibrant area of research to which advanced source separation 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 [39, 40, 41] are also needed to increase the insight that can be gained from evaluation. Finally, increased publicity and networking 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. As the first trial in this direction, the biomedical part of SiSEC 2010 clearly had a limited scope.

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 [23, 34, 27, 42], 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] M. S. Brandstein, D. B. Ward (Eds.), Microphone Arrays: Signal Processing Techniques and Applications, Springer, 2001.

[2] S. Makino, T.-W. Lee, H. Sawada (Eds.), Blind speech separation, Springer, 2007.

[3] P. Comon, C. Jutten (Eds.), Handbook of Blind Source Separation, Independent Component Analysis and Applications, Academic Press, 2010.

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

[5] X. Jaureguiberry, P. Leveau, S. Maller, J. J. Burred, Adaptation of source-specific dictionaries in non-negative matrix factorization for source separation, in: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2011, pp. 5–8.

[6] E. Vincent, H. Sawada, P. Bofill, S. Makino, 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), 2007, pp. 552–559.

[7] E. Vincent, S. Araki, 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), 2009, pp. 734–741.

[8] S. Araki, A. Ozerov, V. Gowreesunker, H. Sawada, F. Theis, G. Nolte, D. Lutter, 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), 2010, pp. 114–122.

[9] S. Araki, F. Theis, G. Nolte, D. Lutter, A. Ozerov, V. Gowreesunker, H. Sawada, 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), 2010, pp. 123–130.

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

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

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

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

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

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

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

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

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

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

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

[21] 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), 2007, pp. 430–437.

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

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

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

[25] H. Sawada, S. Araki, 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) (2011) 516–527.

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

[27] F. Nesta, M. Omologo, 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), 2008, pp. 43–48.

[28] H. Sawada, S. Araki, 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), 2007, pp. 3247–3250.

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

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

[31] N. Q. K. Duong, E. Vincent, 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), 2010, pp. 73–80.

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

[33] S.-I. Lee, S. Batzoglou, Application of independent component analysis to microarrays, Genome Biology 4 (2003) R76.

[34] F. Blöchl, A. Kowarsch, 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), 2010, pp. 434–441.

[35] W. Chen, C. Q. Chang, 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), 2010, pp. 1061–1064.

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

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

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

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

[40] V. Emiya, E. Vincent, 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), 2009, pp. 97–100.

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

[42] V. Y. F. Tan, C. Févotte, Automatic relevance determination in nonnegative matrix factorization, in: Proc. Workshop on Signal Processing with Adaptative Sparse Structured Representations (SPARS), 2009.