First stereo audio source separation evaluation campaign: data, algorithms and results

Emmanuel Vincent, Hiroshi Sawada, Pau Bofill, Shoji Makino, Justinian Rosca

First Stereo Audio Source Separation Evaluation Campaign: Data, Algorithms and Results

Emmanuel Vincent[1], Hiroshi Sawada[2], Pau Bofill[3], Shoji Makino[2], and Justinian P. Rosca[4]

[1] METISS Group, IRISA-INRIA
Campus de Beaulieu, 35042 Rennes Cedex, France
emmanuel.vincent@irisa.fr
[2] Signal Processing Research Group, NTT Communication Science Labs
2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
[3] Departament d'Arquitectura de Computadors, Universitat Politècnica de Catalunya
Campus Nord Mòdul D6, Jordi Girona 1-3, 08034 Barcelona, Spain
[4] Siemens Corporate Research
755 College Road East, Princeton NJ 08540, USA

Abstract. This article provides an overview of the first stereo audio source separation evaluation campaign, organized by the authors. Fifteen underdetermined stereo source separation algorithms have been applied to various audio data, including instantaneous, convolutive and real mixtures of speech or music sources. The data and the algorithms are presented and the estimated source signals are compared to reference signals using several objective performance criteria.

1 Introduction

Large-scale evaluations facilitate progress in a field by revealing the effects of different choices in algorithm design, promoting common test data and evaluation criteria and attracting the interest of funding bodies. Several evaluations of audio source separation algorithms have been conducted recently, focusing on single-channel speech mixtures[5] or multichannel over-determined speech mixtures[6][7][8]. This article provides an overview of the complementary evaluation campaign for stereo underdetermined audio mixtures organized by the authors. Detailed results of the campaign are available at http://sassec.gforge.inria.fr/.

We define the source separation task and describe test data and evaluation criteria in Section 2. Then we present the algorithms submitted by the participants in Section 3 and summarize their results in Section 4. We conclude in Section 5.

[5] http://www.dcs.shef.ac.uk/~martin/SpeechSeparationChallenge.htm
[6] http://bme.engr.ccny.cuny.edu/faculty/parra/bss/
[7] http://homepages.inf.ed.ac.uk/mlincol1/SSC2/
[8] http://mlsp2007.conwiz.dk/index.php?id=43

2 Data and evaluation criteria

2.1 The stereo underdetermined source separation task

Common audio signals, *e.g.* radio, television, music CDs and MP3s, are typically available in *stereo* (two-channel) format and consist of a mixture of more than two sound sources. Denoting by $J > 2$ the number of sources, each channel $x_i(t)$ $(1 \leq i \leq 2)$ of the mixture signal can be expressed as [1]

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

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.

Different types of mixtures can be distinguished. *Instantaneous* mixtures are generated via (1) using a mixing desk or dedicated software by constraining the spatial images of each source j to $s_{ij}^{\text{img}}(t) = a_{ij} s_j(t)$, where $s_j(t)$ is a single-channel source signal and a_{ij} are positive mixing gains. Synthetic *convolutive* mixtures are obtained similarly via $s_{ij}^{\text{img}}(t) = \sum_\tau a_{ij}(\tau) s_j(t-\tau)$, where $a_{ij}(\tau)$ are mixing filters. *Live recordings* are acquired by recording all the sources simultaneously in a room using a pair of microphones. These recordings may also be obtained by recording the sources one at a time in the same room and adding the resulting source images together within each channel [2].

We define the source separation task as that of estimating the spatial images $s_{ij}^{\text{img}}(t)$ of all sources j on all channels i from the two channels $x_i(t)$ of a mixture. This definition has two advantages: it is valid for all types of mixtures, even with spatially extended sources that cannot be represented as single-channel signals, and potential gain or filtering indeterminacies about the estimated single-channel source signals $s_j(t)$ disappear when considering their spatial images instead [1].

2.2 Development and test data

The development and test data used for the evaluation campaign involved four classes of signals: male speech, female speech, non-percussive music and music including drums. Music mixtures involved three sources taken from synchronized multitrack recordings, while speech mixtures involved four independent sources. All the source signals were sampled at 16 kHz and had a duration of 10 s.

The development data consisted of one instantaneous mixture, two synthetic convolutive mixtures and two live recordings per class. Instantaneous mixtures were generated by scaling the source signals by positive gains. Live recordings were acquired by playing the source signals through loudspeakers in a room at NTT with $\text{RT}_{60} = 250\,\text{ms}$ reverberation time and recording them using two pairs of omnidirectional microphones with spacings of 5 cm and 1 m. Figure 1 depicts the arrangement of loudspeakers and microphones. Synthetic convolutive mixtures were obtained by filtering the sources with simulated room impulse responses computed for the same arrangement using Roomsim[9]. Ground truth

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

data, *i.e.* the source signals, their spatial images and the mixing filters or gains, were distributed with the mixture signals at `http://sassec.gforge.inria.fr/`.

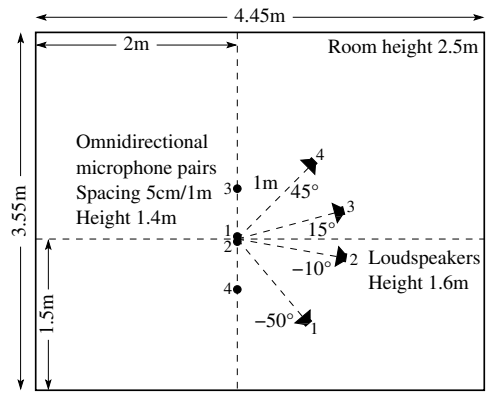

Fig. 1. Recording arrangement used for development data. Only three of the four loudspeakers were used for music mixtures.

The same number of test data was obtained similarly to the development data, using different source signals and positions for each mixture. The distances of the sources from the center of the microphone pairs were drawn randomly between 80 cm and 1.2 m and their angles of arrival between $-60°$ and $+60°$ with a minimal spacing of $15°$. The mixture signals were made available, but ground truth data, including the exact source positions, was kept hidden[10].

2.3 Objective performance criteria

The participants were asked to provide estimates $\hat{s}_{ij}^{\text{img}}(t)$ of the spatial images of all sources j for some test mixtures. The quality of these estimates was then evaluated by comparison with the true source images $s_{ij}^{\text{img}}(t)$ using four objective performance criteria, inspired from criteria previously designed for single-channel source estimates [3]. By contrast with other existing measures [4,5], the proposed criteria can be computed for all types of separation algorithms and do not necessitate knowledge of the separating filters or masks.

The criteria derive from the decomposition of an estimated source image as

$$\hat{s}_{ij}^{\text{img}}(t) = s_{ij}^{\text{img}}(t) + e_{ij}^{\text{spat}}(t) + e_{ij}^{\text{interf}}(t) + e_{ij}^{\text{artif}}(t) \tag{2}$$

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 auditory distinction between sounds from the target source, sounds from other sources and "gurgling"

[10] Only the first two authors of this article had potentially access to these data.

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. The computational modeling of this auditory segregation process is an open issue so far. For simplicity, we chose to express spatial distortion and interference components as filtered versions of the true source images, computed by least-squares projection of the estimated source image onto the corresponding signal subspaces [3]

$$e_{ij}^{\text{spat}}(t) = P_j^L(\hat{s}_{ij}^{\text{img}})(t) - s_{ij}^{\text{img}}(t) \tag{3}$$

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

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

where P_j^L is the least-squares projector onto the subspace spanned by $s_{kj}^{\text{img}}(t-\tau)$, $1 \leq k \leq I, 0 \leq \tau \leq L-1$, and 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 filter length L was set to 512 (32 ms), which was the maximal tractable length.

The relative amounts of spatial distortion, interference and artifacts were then measured using three energy ratio criteria expressed in decibels (dB): the source Image to Spatial distortion Ratio (ISR), the Source to Interference Ratio (SIR) and the Sources to Artifacts Ratio (SAR), defined by

$$\text{ISR}_j = 10\log_{10}\frac{\sum_{i=1}^I \sum_t s_{ij}^{\text{img}}(t)^2}{\sum_{i=1}^I \sum_t e_{ij}^{\text{spat}}(t)^2} \tag{6}$$

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

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

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

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

We emphasize that this measure is arbitrary, in the sense that it weights the three error components equally. In practice, each component should be given a different weight depending on the application. For instance, spatial distortion is of little importance for most applications, except for karaoke where it can result in imperfect source cancellation, even in the absence of interference or artifacts. Similarly, artifacts are crucial for hearing aid applications, for which "gurgling" noise should be avoided at the cost of increased interference. These criteria were implemented in Matlab and distributed at `http://sassec.gforge.inria.fr/`.

3 Algorithms

The campaign involved thirteen participants, who submitted the results of fifteen source separation algorithms. The underlying approaches are summarized in Ta-

Table 1. Submitted source separation algorithms.

N°	Submitter Name	Source localization	Source signal estimation
Algorithms for instantaneous mixtures only			
1	D. Barry ADRess	Manual IID clustering from a magnitude-weighted histogram with auditory feedback [6]	Source magnitude estimation in the STFT bins associated with each IID cluster [6]
2	P. Bofill	Peak picking on a smoothed IID histogram [7] with STFT bins selected as in [8]	Minimization of the l_1 norm of the real and imaginary parts of the source STFTs [9]
3	A. Ehmann	Manual peak picking on an IID histogram	Binary STFT masking with different resolutions at high/low frequencies
4	V. Gowreesunker	Peak picking on a thresholded IID histogram [10]	Binwise MDCT projection onto the nearest IID subspace [10]
5	M. Kleffner	Peak picking on a thresholded IID histogram [11] with STFT bins selected as in [12]	Online FFT-domain minimum-variance beamforming [13]
6	N. Mitianoudis	Soft IID clustering given the number of sources [14]	Binwise MDCT projection onto the nearest IID subspace [14]
7	H. Sawada	Hard IID clustering given the number of sources	Binary STFT masking
8	E. Vincent	Manual peak picking on an IID histogram weighted as in [12]	Minimization of the l_0 norm of the source STFTs [15]
9	M. Xiao SABM+SSDP	Hard fixed-width IID clustering on selected STFT bins [8]	Mixing inversion with 2 sources per time frame estimated from the mixture covariance [16]
10	M. Xiao SABM+SNSDP	Hard fixed-width IID clustering on selected STFT bins [8]	Extension of [16] with more active sources in some time frames
Algorithms for instantaneous and/or convolutive mixtures			
11	S. Araki	Soft (IID,ITD) clustering given the number of sources [17]	Maximum SNR beamforming [18] and soft STFT masking [19]
12	Y. Izumi	Soft clustering of the mixture STFT bins based on (IID,IPD) given the number of sources [20]	Soft STFT masking by cluster probabilities [20]
13	T. Kim		FFT-domain independent component analysis [21] and soft masking (two sources only)
14	R. Weiss & M. Mandel	Soft (IID,IPD) clustering given the number of sources [22]	Soft STFT masking by cluster probabilities [22]
15	H. Sawada	Frequency-wise (IID,IPD) clustering given the number of sources as in [17] and sorting [23]	Binary STFT masking

ble 1. All algorithms except n°13 could be broken into (possibly iterated) source localization and source signal estimation steps. These two steps were conducted in the time-frequency domain via a Short-Time Fourier Transform (STFT) or a

Table 2. Results for instantaneous mixtures.

Algorithm	1	2	3	4	5[11]	6	7	8	9	10	14
SDR (dB)	4.0	4.2	6.8	3.5	-23.4	-16.0	7.2	10.3	5.8	2.7	-2.4
ISR (dB)	7.5	8.2	13.9	6.2	-21.8	-12.8	14.6	19.2	15.9	20.0	4.1
SIR (dB)	13.2	12.9	15.5	14.4	12.8	13.2	15.9	16.0	10.7	6.8	-3.0
SAR (dB)	5.3	10.8	7.8	5.5	5.9	5.3	8.1	12.2	5.8	8.7	4.2
Time (s)	1	300	5	10	600	200	9	5	2	2	1000

Table 3. Results for synthetic convolutive mixtures and live recordings with two different microphone spacings.

Mixtures	Synth 5 cm			Synth 1 m		Live 5 cm					Live 1 m		
Algorithm	11[11]	14	15	14	15	11[11]	12[11]	13[12]	14	15	13[12]	14	15
SDR (dB)	2.5	0.9	0.2	0.7	0.6	2.6	-23.2	-20.3	1.2	1.8	-19.0	2.1	3.6
ISR (dB)	6.0	2.8	4.6	2.8	4.4	5.9	-19.2	-17.0	4.0	7.0	-15.5	4.9	8.4
SIR (dB)	5.8	-2.7	4.4	-0.4	4.2	4.6	1.3	2.9	-1.9	4.2	2.9	0.8	6.9
SAR (dB)	4.9	14.1	7.5	10.7	7.5	5.4	6.2	6.2	13.0	6.8	5.8	8.0	6.8
Time (min)	1	20	0.6	20	0.6	1	1	4	20	0.6	4	20	0.6

Modified Discrete Cosine Transform (MDCT), except for algorithms n°9 and 10 where source estimation was directly performed in the time domain. The directions of the sources were modeled by the Interchannel Intensity Difference (IID) or variants thereof in the instantaneous case. The Interchannel Time Difference (ITD) or the Interchannel Phase Difference (IPD) were additionally used in the convolutive case. Algorithms n°2, 4, 5, 9 and 10 were fully blind, while others required manual input of the number of sources or the source directions.

4 Results

The performance of each algorithm was assessed by sorting the estimated source image signals so as to maximize the average SIR and successively averaging the measured SDR, ISR, SIR and SAR over the sources and over the mixtures. The resulting figures are given in Tables 2 and 3 for instantaneous and convolutive mixtures respectively, along with platform-specific computation times. The large negative SDR and ISR figures for algorithms n°5, 6, 12 and 13 are due to incorrect scaling of the submitted source images. Detailed results and sound files are available at `http://sassec.gforge.inria.fr/`.

In the instantaneous case, most algorithms provided similar SIR and SAR values clustered around 13 dB and 6 dB respectively, denoting high interference rejection but clear artifacts. Algorithms n°2 and 8 resulted in fewer artifacts, while algorithms n°10 and 14 provided more interference. Note that blind algorithms n°9 and 10 achieved similar source localization accuracy as non-blind algorithms n°3, 7 and 8, as shown by large ISR values.

[11] Average performance for speech mixtures only.
[12] Average performance over the two estimated sources for speech mixtures only.

In the convolutive case, most algorithms provided again similar SIR and SAR values but around 4 dB and 6 dB respectively, indicating both strong interference and artifacts. Algorithms n°11 and 15 resulted in slightly less interference, while algorithm n°14 provided much more interference but less artifacts. Interestingly, performance did not vary much between synthetic convolutive mixtures and live recordings or with different microphone spacings.

5 Conclusion

In this article, we described the test data and objective performance criteria used in the context of the first stereo audio source separation evaluation campaign and summarized the approaches behind the fifteen submitted algorithms and their results. We are currently planning to complement objective performance figures by listening tests and present detailed results on the campaign website. We hope that this campaign fosters interest for evaluation in the source separation community and that larger-scale regular campaigns will take place in the future. The creation of a collaborative organization framework appears crucial to this aim, since it would allow sharing between the participants of time-consuming tasks such as the collection of test data under appropriate licenses and the recording of live mixtures.

References

1. Cardoso, J.F.: Multidimensional independent component analysis. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP). (1998) IV–1941–1944
2. Schobben, D., Torkkola, K., Smaragdis, P.: Evaluation of blind signal separation methods. In: Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA). (1999) 261–266
3. Vincent, E., Gribonval, R., Févotte, C.: Performance measurement in blind audio source separation. IEEE Trans. on Audio, Speech and Language Processing **14** (2006) 1462–1469
4. Mansour, A., Kawamoto, M., Ohnishi, N.: A survey of the performance indexes of ICA algorithms. In: Proc. IASTED Int. Conf. on Modelling, Identification and Control (MIC). (2002) 660–666
5. Yılmaz, O., Rickard, S.T.: Blind separation of speech mixtures via time-frequency masking. IEEE Trans. on Signal Processing **52** (2004) 1830–1847
6. Barry, D., Coyle, E., Lawlor, B.: Real-time sound source separation using azimuth discrimination and resynthesis. In: Proc. 117th AES Convention. (2004) preprint 6258.
7. Bofill, P., Zibulevsky, M.: Underdetermined blind source separation using sparse representations. Signal Processing **81** (2001) 2353–2362
8. Xiao, M., Xie, S., Fu, Y.: A novel approach for underdetermined blind source separation in the frequency domain. In: Proc. Int. Symp. on Neural Networks (ISNN). (2005) 484–489

9. Bofill, P., Monte, E.: Underdetermined convoluted source reconstruction using LP and SOCP, and a neural approximator of the optimizer. In: Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA). (2006) 569–576
10. Gowreesunker, B.V., Tewfik, A.H.: Two improved sparse decomposition methods for blind source separation. In: Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA). (2007)
11. Mohan, S., Kramer, M.L., Wheeler, B.C., Jones, D.L.: Localization of nonstationary sources using a coherence test. In: Proc. IEEE Workshop on Statistical Signal Processing (SSP). (2003) 470–473
12. Arberet, S., Gribonval, R., Bimbot, F.: A robust method to count and locate audio sources in a stereophonic linear instantaneous mixture. In: Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA). (2006) 536–543
13. Lockwood, M.E., Jones, D.L., Bilger, R.C., Lansing, C.R., O'Brien Jr., W.D., Wheeler, B.C., Feng, A.S.: Performance of time- and frequency-domain binaural beamformers based on recorded signals from real rooms. Journal of the Acoustical Society of America **115** (2004) 379–391
14. Mitianoudis, N., Stathaki, T.: Underdetermined source separation using mixtures of warped Laplacians. In: Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA). (2007)
15. Vincent, E.: Complex nonconvex l_p norm minimization for underdetermined source separation. In: Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA). (2007)
16. Xiao, M., Xie, S., Fu, Y.: A statistically sparse decomposition principle for underdetermined blind source separation. In: Proc. Int. Symp. on Intelligent Signal Processing and Communication Systems (ISPACS). (2005) 165–168
17. O'Grady, P.D., Pearlmutter, B.A.: Soft-LOST: EM on a mixture of oriented lines. In: Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA). (2004) 428–435
18. Araki, S., Sawada, H., Makino, S.: Blind speech separation in a meeting situation with maximum SNR beamformers. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP). (2007) I–41–44
19. Cermak, J., Araki, S., Sawada, H., Makino, S.: Blind source separation based on beamformer array and time-frequency binary masking. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP). (2007) I–145–148
20. Izumi, Y., Ono, N., Sagayama, S.: Sparseness-based 2ch BSS using the EM algorithm in reverberant environment (2007) Submitted to IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA).
21. Kim, T., Attias, H.T., Lee, S.Y., Lee, T.W.: Blind source separation exploiting higher-order frequency dependencies. IEEE Trans. on Audio, Speech and Language Processing **15** (2007) 70–79
22. Mandel, M.I., Ellis, D.P.W., Jebara, T.: An EM algorithm for localizing multiple sound sources in reverberant environments. In: Advances in Neural Information Processing Systems (NIPS 19). (2007)
23. Sawada, H., Araki, S., Makino, S.: Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS. In: Proc. IEEE Int. Symp. on Circuits and Systems (ISCAS). (2007) 3247–3250