The 2008 Signal Separation Evaluation Campaign: A community-based approach to large-scale evaluation

Emmanuel Vincent, Shoko Araki, Pau Bofill

To cite this version:

Emmanuel Vincent, Shoko Araki, Pau Bofill. The 2008 Signal Separation Evaluation Campaign: A community-based approach to large-scale evaluation. 8th Int. Conf. on Independent Component Analysis and Signal Separation (ICA), Mar 2009, Paraty, Brazil. pp.734–741. inria-00544168

HAL Id: inria-00544168

https://hal.inria.fr/inria-00544168

Submitted on 7 Dec 2010

The 2008 Signal Separation Evaluation Campaign: A Community-Based Approach to Large-Scale Evaluation

Emmanuel Vincent[1], Shoko Araki[2], and Pau Bofill[3]

[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

Abstract. This paper introduces the first community-based Signal Separation Evaluation Campaign (SiSEC 2008), coordinated by the authors. This initiative aims to evaluate source separation systems following specifications agreed between the entrants. Four speech and music datasets were contributed, including synthetic mixtures as well as microphone recordings and professional mixtures. The source separation problem was split into four tasks, each evaluated via different objective performance criteria. We provide an overview of these datasets, tasks and criteria, summarize the results achieved by the submitted systems and discuss organization strategies for future campaigns.

1 Introduction

Large-scale evaluations are a key ingredient to scientific and technological maturation by revealing the effects of different system designs, promoting common test specifications and attracting the interest of industries and funding bodies. Recent evaluations of source separation systems include the 2006 Speech Separation Challenge[4] and the 2007 Stereo Audio Source Separation Evaluation Campaign [1]. The subsequent panel discussion held at the 7th International Conference on Independent Component Analysis and Signal Separation (ICA 2007) resulted in a set of recommendations regarding future evaluations, in particular:

- splitting the overall problem into several successive or alternative tasks,
- providing reference software and evaluation criteria for each task,
- considering toy data as well as real-world data of interest to companies,
- letting entrants specify all aspects of the evaluation collaboratively.

These general principles aim to facilitate the entrance of researchers addressing different tasks and to enable detailed diagnosis of the submitted systems.

[4] http://www.dcs.shef.ac.uk/~martin/SpeechSeparationChallenge.htm

This article introduces the 2008 community-based Signal Separation Evaluation Campaign (SiSEC) as a tentative implementation of these principles. Due to the variety of the submitted systems, we focus on the general outcomes of the campaign and let readers refer to the website at `http://sisec.wiki.irisa.fr/` for the details and results of individual systems. We describe the chosen datasets, tasks and evaluation criteria in Section 2. We summarize the results and provide bibliographical references to the submitted systems in Section 3. We conclude and discuss organization strategies for future campaigns in Section 4.

2 Specifications

The datasets, tasks and evaluation criteria considered in the campaign were specified in a collaborative fashion. A few initial specifications were first suggested 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. All materials, including data and code, are available at `http://sisec.wiki.irisa.fr/`.

2.1 Datasets

The data consisted of audio signals spanning a range of mixing conditions. The channels $x_i(t)$ $(1 \leq i \leq I)$ of each mixture signal were generally obtained as

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

where $s_{ij}^{\mathrm{img}}(t)$ is the *spatial image* of source j $(1 \leq j \leq J)$ on channel i, that is the contribution of this source to the mixture in this channel. *Instantaneous* mixtures are generated via $s_{ij}^{\mathrm{img}}(t) = a_{ij} s_j(t)$, where $s_j(t)$ are single-channel source signals and a_{ij} positive mixing gains. *Convolutive* mixtures are obtained similarly from mixing filters $a_{ij}(\tau)$ via $s_{ij}^{\mathrm{img}}(t) = \sum_\tau a_{ij}(\tau) s_j(t-\tau)$. *Recorded* mixtures are acquired by playing each source at a time on a loudspeaker and recording it over a set of microphones. Four distinct datasets were provided:

D1 Under-determined speech and music mixtures
This dataset consists of 36 instantaneous, convolutive and recorded stereo mixtures of three to four audio sources of 10 s duration, sampled at 16 kHz. Recorded mixtures were acquired in a chamber with cushion walls, using the loudspeaker and microphone arrangement depicted in [1], while convolutive mixtures were obtained with artificial room impulse responses simulating the same arrangement. The distance between microphones was set to either 5 cm or 1 m and the room reverberation time (RT) to 130 ms or 250 ms. The source signals include unrelated female or male speech and synchronized percussive or non-percussive music.

D2 Determined and over-determined speech and music mixtures
This dataset includes 21 four-channel recordings of two to four unrelated speech sources of 10 s duration, sampled at 16 kHz, acquired in four different rooms: two chambers with cushion walls, an office room and a conference room. Some mixtures were directly recorded instead of computed via (1). The microphones were placed either at a height of 1.25 m or at different heights, near the walls or near the center and at a distance of about 5 cm or 1 m. The sources were placed either randomly or at 1 m distance from the microphones. The dataset also includes a 2-channel mixture of 2 speech sources recorded via cardioid microphones placed on either side of a dummy head.

D3 Head-geometry mixtures of two speech sources in real environments
This dataset consists of 648 two-channel convolutive mixtures of two unrelated speech sources of about 10 s, sampled at 16 kHz. The mixing filters were real-world impulse responses from two rooms, an anechoic chamber and an office room, measured by hearing aid microphones mounted on a dummy head. The sources were placed in the horizontal plane at fixed distance from the head. In the anechoic chamber, the distance was set to 3 m and the direction of arrival (DOA) varied over 360° in 20° steps. In the office room, the distance was set to 1 m and the DOA varied over the front 180° hemisphere in 10° steps. All possible combinations of two different DOAs were generated.

D4 Professionally produced music recordings
This dataset consists of two stereo music signals sampled at 44.1 kHz involving two and ten synchronized sources of 13 and 14 s duration, respectively. The stereo spatial image of each source was generated by a combination of professional recording and mixing techniques. Special effects applied to individual sources include chorus, distortion pads, vocoder, delays, parametic equalization and dynamic multi-band compression.

All datasets except D2 include both test and development data generated in a similar fashion, but from different source signals and source positions. The true source signals and source positions underlying the test data were hidden to the entrants, except for D3 where the source positions were provided as prior information. The true number of sources was always available.

2.2 Tasks

The source separation problem was split into four tasks:

T1 Source counting
T2 Mixing system estimation
T3 Source signal estimation
T4 Source spatial image estimation

These tasks consists of finding, respectively: (T1) the number of sources J, (T2) the mixing gains a_{ij} or the discrete Fourier transform $a_{ij}(\nu)$ of the mixing filters, (T3) the single-channel source signals $s_j(t)$ and (T4) the spatial images $s_{ij}^{\mathrm{img}}(t)$ of the sources over all channels i. Entrants were asked to submit the results of their system to T3 and/or T4 and on an optional basis to T1 and/or T2.

Reference software was provided to address tasks T1 and T2 over instantaneous mixtures [2] (R1) and tasks T3 and T4 either via binary masking (R2) or via l_p-norm minimization [3] (R3). This software aims to facilitate entrance and to provide baseline results for benchmarking purposes. Two oracle systems were also considered for the benchmarking of task T4: ideal binary masking over a short-time Fourier transform (STFT) [4] (O1) or over a cochleagram [5] (O2). These systems require knowledge of the true source spatial images and provide theoretical upper performance bounds for binary masking-based systems.

2.3 Evaluation criteria

Although standard evaluation criteria exist for task T2 when the number of sources is smaller than the number of sensors, there are no such criteria in a more general setting so far. For instantaneous mixtures, the vector $\widehat{\mathbf{a}}_j$ of estimated mixing gains for a given source j was decomposed as

$$\widehat{\mathbf{a}}_j = \mathbf{a}_j^{\mathrm{coll}} + \mathbf{a}_j^{\mathrm{orth}} \tag{2}$$

where $\mathbf{a}_j^{\mathrm{coll}}$ and $\mathbf{a}_j^{\mathrm{orth}}$ are respectively collinear and orthogonal to the true vector of mixing gains $\mathbf{a}_j$ and are computed by least squares projection. Accuracy was then assessed via the mixing error ratio (MER) in decibels (dB)

$$\mathrm{MER}_j = 10 \log_{10} \frac{\|\mathbf{a}_j^{\mathrm{coll}}\|^2}{\|\mathbf{a}_j^{\mathrm{orth}}\|^2} \tag{3}$$

where $\|.\|$ is the Euclidean norm. This criterion allows arbitrary scaling of the gains for each source. It is equal to $+\infty$ for an exact estimate, 0 when the estimate forms a $45°$ angle with the ground truth and $-\infty$ when it is orthogonal. For convolutive mixtures, the accuracy of estimated mixing filters for source j was similarly assessed by computing the MER in each frequency bin ν between $\widehat{\mathbf{a}}_j(\nu)$ and $\mathbf{a}_j(\nu)$ and averaging it over frequency. Since the sources can be characterized only up to an arbitrary permutation, all possible permutations were tested and the one maximizing the average MER was selected.

Tasks T3 and T4 were evaluated via the criteria in [6] and [1], respectively, termed signal to distortion ratio (SDR), source image to spatial distortion ratio (ISR), signal to interference ratio (SIR) and signal to artifacts ratio (SAR). These criteria can be computed for any separation system and do not necessitate knowledge of the unmixing filters or masks. The SDR for task T3 allows arbitrary filtering of the target source, while that for T4 allows no scaling or filtering distortion, which is separately measured by the ISR. The signals were permuted so as to maximize the average SIR. The resulting permutations were found to be relevant and identical to that estimated from the MER, except in cases involving much interference. For dataset D4, performance was also measured via a magnitude Signal-to-Error Ratio (mSER) between the true and estimated magnitude STFTs of the source spatial images over each channel.

Table 1. Average MER for task T2 over the instantaneous mixtures of dataset D1.

System	[11]	[7]	[8]	R1
MER	80.9	81.7	42.4	49.0

Table 2. Average performance for tasks T3 or T4 over the instantaneous mixtures of dataset D1. Figures relate to T4 when the ISR is reported and to T3 otherwise.

System	[12]	[13]	[3]	[14]	[15]	[11][5]	[16]	[17]	[9][5]	[8]	R2	R3	O1	02
SDR	9.8	14.0	11.7	11.3	7.8	10.7	12.6	6.8	6.2[6]	5.5	8.8	11.1	10.7	8.1
ISR	18.2	24.1	22.9	20.1	16.4		21.9			11.4	17.5	22.7	20.0	14.4
SIR	14.9	20.6	18.5	16.9	17.3	18.6	18.5	16.8	11.6[6]	12.9	18.6	18.4	21.6	17.4
SAR	11.4	15.4	13.2	12.8	8.8	11.9	13.2	7.7	9.0[6]	8.2	9.6	12.3	11.5	9.1

Table 3. Average performance for tasks T3 or T4 over the convolutive/recorded mixtures of dataset D1. Figures relate to T4 when the ISR is reported and to T3 otherwise.

System	RT=130ms				RT=250ms			
	SDR	ISR	SIR	SAR	SDR	ISR	SIR	SAR
[18]	2.0[6]	6.3[6]	5.8[6]	5.5[6]	0.9[6]	5.1[6]	2.8[6]	5.1[6]
[19]	1.9	4.8	2.7	5.7	1.2	4.1	0.6	6.1
[20]	2.2	4.2	3.2	7.1	1.5	3.5	1.6	7.9
[10]	2.9[6]	6.5[6]	7.1[6]	8.6[6]	3.7[6]	6.5[6]	5.0[6]	8.8[6]
[21]	-1.1[6]		6.8[6]	1.3[6]	-1.1[6]		6.6[6]	1.5[6]
[8]	3.3	6.7	4.3	7.9	3.1	6.2	3.9	8.4
O1	9.7	18.3	19.9	10.2	8.7	16.2	19.4	10.4
O2	6.9	12.1	16.5	7.6	6.6	11.5	16.0	7.9

3 Results

The details and the results of the thirty submitted systems are available for viewing and listening at `http://sisec.wiki.irisa.fr/`. The systems [7], [8], R1, [9] and [10] addressed task T1 without error. Summary performance figures for other tasks are provided in Tables 1 to 6 and in Figure 1 after averaging over all sources then over several mixtures. An analysis of each table is beyond the scope of this paper, due to the wide variety of prior knowledge and computation resources used by different systems. We observe that the mixing matrix estimation task is now solved for instantaneous mixtures, that the source signal estimation task can now be addressed with a mean SIR around 20 dB for instantaneous or anechoic mixtures and that the separation of monophonic instruments from professional music recordings can also be achieved with a SIR above 15 dB. Nevertheless, the separation of reverberant mixtures remains a challenge for any number of sources and channels despite continued progress, as illustrated by an average SIR around 6 dB for office recordings of three sources.

[5] Variant or extension of the system presented in the bibliographical reference.

[6] Figure computed by averaging over an incomplete set of mixtures or sources.

Table 4. Average SIR for task T3 over dataset D2.

System	Cushioned rooms			Office/lab rooms			Conference room		
	$J = 2$	$J = 3$	$J = 4$	$J = 2$	$J = 3$	$J = 4$	$J = 2$	$J = 3$	$J = 4$
[22]	14.4	16.3	8.9	14.1	5.7	-0.3	8.2	1.5	-2.3
[23]	5.3	12.8	9.0	19.6[6]					
[24]	3.9	4.3	0.9	6.9	2.6	-2.9	7.1	2.2	-0.6
I. Takashi	11.3[6]	7.4[6]	5.6[6]	2.8[6]					
[25]	10.6	9.2	4.1	4.2[6]	-0.4	-3.7	2.3	-1.2	-3.5
[26]	8.7	6.8	2.5	3.2[6]	-1.3	-4.0	2.2	-1.3	-5.1

Table 5. Average performance for task T4 over dataset D3.

System	Anechoic chamber				Office room			
	SDR	ISR	SIR	SAR	SDR	ISR	SIR	SAR
[19]	10.8	19.9	19.2	13.6	4.1	8.6	5.8	10.7
[19][5]	11.3	19.9	19.3	14.3	4.4	8.8	5.9	11.2
[27]					3.7	9.2	6.1	10.7
O1	13.7	24.5	24.7	14.1	13.0	23.7	23.9	13.4
O2	11.0	20.1	19.9	11.7	10.6	19.8	19.6	11.3

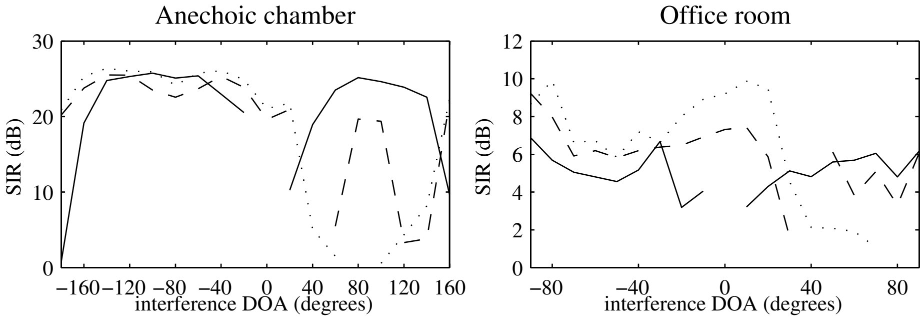

Fig. 1. Average SIR achieved by system [19] for task T4 over dataset D3 as a function of the interference DOA for three target DOAs (plain: 0°, dashed:40°, dotted:80°).

4 Conclusion

We summarized the specifications and outcomes of the first community-based Signal Separation Evaluation Campaign. We hope that this campaign fosters interest for evaluation in the source separation community, so that more entrants contribute feedback, datasets or code in the future. With thirty submissions but three organizers only, the current organization scheme has reached its goal of attracting many entrants, but failed to provide detailed analysis of the results. We believe that increased participation from the community is key to maximizing the benefits of future campaigns. We advocate the creation of a larger organization committee with members dedicated to the evaluation of a particular dataset or task and invite all willing researchers to become part of it.

Table 6. Average performance for task T4 over dataset D4. The SIR quantifies interference from the target sources only, while the SAR includes that from other sources.

System	*Tamy* ($J = 2$)					*Bearlin* ($J = 10$)				
	mSER	SDR	ISR	SIR	SAR	mSER	SDR	ISR	SIR	SAR
[13]	6.0	4.5	10.0	8.9	8.6	4.2	-0.4	7.8	6.9	1.6
[16][5]	6.8	5.9	10.2	8.8	10.7	4.9	3.8	9.7	8.3	4.8
[28]	9.5	8.6	17.3	16.4	9.5					
[29]	8.4	7.7	16.5	15.4	8.4					
[15][5]	8.5	7.2	16.5	15.7	8.3	3.3	2.6	8.6	12.9	1.6
[30]	3.5	4.5	7.4	18.5	4.6	3.4	3.2	8.4	12.0	1.8
[31][5]	9.3	8.3	15.1	23.5	8.0					
[31][5]	10.0	9.1	15.1	24.1	9.1					
[31][5]						5.7[6]	5.4[6]	9.8[6]	17.7[6]	5.2[6]
O1	12.8	11.0	21.4	21.1	11.4	9.1	7.5	14.7	18.0	7.9
O2	9.0	8.0	14.5	15.1	8.9	-0.3	2.0	8.0	10.7	0.4

Acknowledgments

We would like to thank all the entrants, as well as H. Sawada, M. Dyrholm, L.C. Parra, K.E. Hild II, H. Kayser, J. Anemüller, M. Vinyes Raso and J. Woodruff for sharing their datasets and evaluation code. The collection of dataset D3 was supported by the European Commission under IP DIRAC.

References

1. Vincent, E., Sawada, H., Bofill, P., Makino, S., Rosca, J.: First stereo audio source separation evaluation campaign: Data, algorithms and results. In: Proc. ICA. (2007) 552–559
2. Bofill, P.: Identifying single source data for mixing matrix estimation in instantaneous blind source separation. In: Proc. ICANN. (2008) 759–767
3. Vincent, E.: Complex nonconvex l_p norm minimization for underdetermined source separation. In: Proc. ICA. (2007) 430–437
4. Vincent, E., Gribonval, R., Plumbley, M.D.: Oracle estimators for the benchmarking of source separation algorithms. Signal Processing **87** (2007) 1933–1950
5. Wang, D.L.: On ideal binary mask as the computational goal of auditory scene analysis. In: Speech Separation by Humans and Machines. Springer, New York, NY (2005)
6. 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
7. Arberet, S., Gribonval, R., Bimbot, F.: A robust method to count and locate audio sources in a stereophonic linear instantaneous mixture. In: Proc. ICA. (2006) 536–543
8. El Chami, Z., Pham, A.D.T., Servière, C., Guerin, A.: A new model based underdetermined source separation. In: Proc. IWAENC. (2008)
9. Gowreesunker, B.V., Tewfik, A.H.: Blind source separation using monochannel overcomplete dictionaries. In: Proc. ICASSP. (2008) 33–36

10. Araki, S., Nakatani, T., Sawada, H., Makino, S.: Stereo source separation and source counting with MAP estimation with Dirichlet prior considering spatial aliasing problem. In: Proc. ICA. (2009)
11. Deville, Y., Puigt, M., Albouy, B.: Time-frequency blind signal separation : extended methods, performance evaluation for speech sources. In: Proc. IJCNN. (2004) 255–260
12. Nesbit, A., Vincent, E., Plumbley, M.D.: Extension of sparse, adaptive signal decompositions to semi-blind audio source separation. In: Proc. ICA. (2009)
13. Ozerov, A., Févotte, C.: Multichannel nonnegative matrix factorization in convolutive mixtures. With application to blind audio source separation. In: Proc. ICASSP. (2009) in press.
14. Vincent, E., Arberet, S., Gribonval, R.: Underdetermined instantaneous audio source separation via local Gaussian modeling. In: Proc. ICA. (2009)
15. Cobos, M., López, J.J.: Stereo audio source separation based on time-frequency masking and multilevel thresholding. Digital Signal Processing **18** (2008) 960–976
16. Arberet, S., Ozerov, A., Gribonval, R., Bimbot, F.: Blind spectral-GMM estimation for underdetermined instantaneous audio source separation. In: Proc. ICA. (2009)
17. Gowreesunker, B.V., Tewfik, A.H.: Two improved sparse decomposition methods for blind source separation. In: Proc. ICA. (2007) 365–372
18. Cobos, M., López, J.J.: Blind separation of underdetermined speech mixtures based on DOA segmentation. IEEE Trans. on Signal Processing (2009) submitted.
19. Mandel, M.I., Ellis, D.P.W.: EM localization and separation using interaural level and phase cues. In: Proc. WASPAA. (2007) 275–278
20. 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)
21. Izumi, Y., Ono, N., Sagayama, S.: Sparseness-based 2ch BSS using the EM algorithm in reverberant environment. In: Proc. WASPAA. (2007) 147–150
22. Nesta, F., Omologo, M., Svaizer, P.: Multiple TDOA estimation by using a state coherence transform for solving the permutation problem in frequency-domain BSS. In: Proc. MLSP. (2008) 43–48
23. Lee, I.: Permutation correction in blind source separation using sliding subband likelihood function. In: Proc. ICA. (2009)
24. Lee, I., Kim, T., Lee, T.W.: Independent vector analysis for blind speech separation. In: Blind speech separation. Springer, Dordrecht, The Netherlands (2007)
25. Gupta, M., Douglas, S.C.: Scaled natural gradient algorithms for instantaneous and convolutive blind source separation. In: Proc. ICASSP. (2007) II–637–II–640
26. Douglas, S.C., Gupta, M., Sawada, H., Makino, S.: Spatio-temporal FastICA algorithms for the blind separation of convolutive mixtures. IEEE Trans. on Audio, Speech and Language Processing **15** (2007) 1511–1520
27. Weiss, R.J., Ellis, D.P.W.: Speech separation using speaker-adapted eigenvoice speech models. Computer Speech and Language (2008) in press.
28. Durrieu, J.L., Richard, G., David, B.: Singer melody extraction in polyphonic signals using source separation methods. In: Proc. ICASSP. (2008) 169–172
29. Durrieu, J.L., Richard, G., David, B.: An iterative approach to monaural musical mixture de-soloing. In: Proc. ICASSP. (2009)
30. Bonada, J., Loscos, A., Vinyes Raso, M.: Demixing commercial music productions via human-assisted time-frequency masking. In: Proc. AES 120th Convention. (2006)
31. Cancela, P.: Tracking melody in polyphonic audio. In: Proc. MIREX. (2008)