The 2015 Signal Separation Evaluation Campaign

Nobutaka Ono[1], Zafar Rafii[2], Daichi Kitamura[3], Nobutaka Ito[4], and Antoine Liutkus[5]

[1] National Institute of Informatics, Japan
[2] Media Technology Lab, Gracenote, Emeryville, USA
[3] SOKENDAI (The Graduate University for Advanced Studies), Japan
[4] NTT Communication Science Laboratories, NTT Corporation, Japan
[5] INRIA, Villers-lès-Nancy, France

Abstract. In this paper, we report the 2015 community-based Signal Separation Evaluation Campaign (SiSEC 2015). This SiSEC consists of four speech and music datasets including two new datasets: "Professionally produced music recordings" and "Asynchronous recordings of speech mixtures". Focusing on them, we overview the campaign specifications such as the tasks, datasets and evaluation criteria. We also summarize the performance of the submitted systems.

1 Introduction

Sharing datasets and evaluating methods with common tasks and criteria has recently become a general and popular methodology to accelerate the development of new technologies. Aiming to evaluate signal separation methods, the Signal Separation Evaluation Campaign (SiSEC) has been held about every one-and-half year in conjunction with the LVA/ICA conference since 2008. The tasks, datasets, and evaluation criteria in the past SiSECs are still available online with the results of the participants. They have been referred to and utilized for comparison and further evaluation by researchers in the source separation community, not limited to the past participants, as shown in Figure 1.

In this fifth SiSEC, two new datasets were added: A new music dataset for a large-scale evaluation was provided in "Professionally produced music recordings" and another new dataset including real recording was provided in "Asynchronous recordings of speech mixtures". For further details, the readers are referred to the web page of SiSEC 2015 at `https://sisec.inria.fr/`. In section 2, we specify the tasks, datasets and evaluation criteria, with a particular focus on these new datasets. Section 3 summarizes the evaluation results.

2 Specifications

SiSEC 2015 focused on the following source separation tasks and datasets.

T1 Single-channel source estimation
T2 Multichannel source image estimation

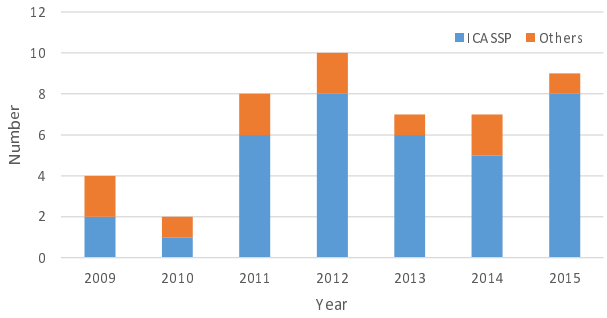

Fig. 1. The number of papers referring SiSEC datasets found by full-text-search on all ICASSP proceedings (ICASSP) and by abstract-search on IEEE Xplore (Others).

D1 Underdetermined speech and music mixtures
D2 Two-channel mixtures of speech and real-world background noise
D3 Professionally produced music recordings
D4 Asynchronous recordings of speech mixtures

T1 aims to estimate single-channel source signals observed by a specific reference microphone, whereas T2 aims to estimate multichannel source images observed by the microphone array.

In D1 and D2, we utilized the same datasets as in SiSEC 2013, which permits easy comparison. Their specifications are given in details in [1].

The new D3 dataset, the Mixing Secret Dataset 100 (MSD100) is designed to evaluate the separation of multiple sources from professionally-produced music recordings. MSD100 consists of 100 full-track songs of different styles, and includes both the stereophonic mixtures and the original stereo sources images. The data is divided into a development set and a test set, each consisting of 50 songs, so that algorithms which need supervised learning can be trained on the development set and tested on the test set. The duration of the songs ranges from 2 minutes and 22 seconds to 7 minutes and 20 seconds, with an average duration of 4 minutes and 10 seconds.

For each song, MSD100 includes 4 stereo sources corresponding to the bass, the drums, the vocals and "other" (i.e., the other instruments). The sources were created using stems from selected raw multitrack projects downloaded from the 'Mixing Secrets' Free Multitrack Download Library[1]. Stems corresponding to a given source were summed together and the result was normalized, then scaled so that the mixture would also be normalized. The mixtures were then generated by summing the sources together. For a given song, the mixture and the sources have the same duration; however, while the mixture is always stereo, some sources can be mono (typically, the vocals). In that case, it appears identical in the left and right channels of the mixture. All items are WAV files sampled at 44.1kHz.

The D4 dataset aims to evaluate the separation of mixtures recorded with asynchronous devices. A new dataset added to D4 contains real recordings of

[1] www.cambridge-mt.com/ms-mtk.htm

three or four speakers using four different stereo IC recorders (8 channels in total). A standard way to make datasets for BSS evaluation is to record each source image first, which is used as the ground truth, and then to make a mixture by summing them up. Unlike conventional synchronized recording, it is not easy in an asynchronous setting because the time offset (time of recording start) of each device is unknown and because there is a sampling frequency mismatch between channels. To obtain consistent source images and real mixtures, a chirp signal was played back from a loudspeaker for time-marking, and the time offsets at the different devices were aligned precisely at a sub-sample level. It is assumed that the sampling frequency of each device is invariant over the whole recording. This dataset consists of three types of mixing: *realmix*, *sumrefs* and *mix*. The realmix is a recording of the real mixture, the sumrefs is the summation of the source images, and the mix is the simulated mixture generated by convolving impulse responses with the dry source and applying resampling for the artificial sampling frequency mismatch.

The BSS Eval toolbox [2] was used to evaluate the following four power-based criteria: the signal to distortion ratio (SDR), the source image to spatial distortion ratio (ISR), the signal to interference ratio (SIR), and signal to artifacts ratio (SAR). The version 2.0 of the PEASS toolbox [3] was used to evaluate the following four perceptually-motivated criteria: the overall perceptual score (OPS), the target-related perceptual score (TPS), the interference-related perceptual score (IPS), and the artifact-related perceptual score (APS). More specifically, T1 was evaluated by `bss_eval_source_denoising.m` for D2 or `bss_eval_source.m` for others. T2 on D3 and D4 was evaluated with `bss_eval_image.m`. For D1 and D2, the PEASS toolbox was used for the comparison with previous SiSEC.

3 Results

We evaluated 27 algorithms in total: 3, 2, 19, and 3 algorithms for D1, D2, D3 and D4, respectively. The average performance of the systems is summarized in Tables 1 to 3, and Figures 2 and 3. Because of the space limitation, only part of the results is shown.

Three algorithms were submitted to D1 as shown in Table 1. Sgouros's method [4] for instantaneous mixtures is based on direction of arrival (DOA) estimation by fitting a mixture of directional Laplacian distributions. The other two algorithms are for convolutive mixtures. Bouafif's method [5] exploits a detection of glottal closure instants in order to estimate the number of speakers and their time delays of arrival (TDOA). It also aims at separation with less artifacts and distortion. Indeed, it shows higher SARs and APSs. However, the SIRs and IPSs are lower. This fact illustrates the well known trade-off between SIR and SAR in BSS. Nguyen's method is similar to [6] and the permutation problem is solved by multi-band alignment [25]. Overall, the performance is almost equivalent to the past SiSEC, which indicates that underdetermined BSS for convolutive mixtures is still a tough problem.

Two algorithms were submitted to D2 as shown in Table 3. López's method [7] designs the demixing matrix and the post-filters based on a single-channel source

separation method. In this submission, they used spectral subtraction as the single-channel source separation method. Note that the performance may vary depending on the choice of the single-channel method. Ito's method is based on full-band clustering of the time-frequency components [8]. Thanks to a frequency-independent time-varying source presence model, the method robustly solves the permutation problem and shows good denoising performance even though it does not explicitly include spectral modeling of speech and noise.

Similarly to the previous SiSEC, D3 attracted most participants. The evaluated methods includes 5 methods available online (not submitted by participants) and are as follows.

- CHA: system using a two-stage Robust Principal Component Analysis (RPCA)[2], with an automatic vocal activity detector and a melody detector [9].
- DUR1, DUR2: systems using a source-filter model for the voice and a Non-negative Matrix Factorization (NMF) model for the accompaniment[3], without (DUR1) and with (DUR2) unvoiced vocals model [10].
- HUA1, HUA2: systems using RPCA[4], with binary (HUA1) and soft (HUA2) masking [11].
- KAM1, KAM2, KAM3: systems using Kernel Additive Modelling (KAM), with light kernel additive modelling (KAM1)[5], a variant with only one iteration (KAM2), and a variant where the energy of the vocals is adjusted at each iteration (KAM3) [12, 13].
- NUG1, NUG2, NUG3: systems using spatial covariance models and Deep Neural Networks (DNN) for the spectrograms, with one set of four DNNs for the four sources for all the iterations (NUG1), one set for the first iteration and another set for the subsequent iterations (NUG2), and one DNN for all the sources (NUG3) [14].
- OZE: system using the Flexible Audio Source Separation Toolbox (FASST) (version 1)[6] [15, 16].
- RAF1, RAF2, RAF3: systems using the REpeating Pattern Extraction Technique (REPET)[7], with the original REPET with segmentation (RAF1), the adaptive REPET (RAF2), and REPET-SIM (RAF3) [17–20].
- STO: system using a predominant pitch extraction and an efficient comb filtering[8] [21, 22].
- UHL1, UHL2, UHL3: systems using DNN, with an independent training material, with four DNNs for the four sources (UHL1), then augmented with an extended training material (UHL2), then using a phase-sensitive cost function (UHL3) [23, 24].
- Ideal: system using the ideal soft masks computed from the mixtures and the sources.

[2] http://mac.citi.sinica.edu.tw/ikala/
[3] http://www.durrieu.ch/research/jstsp2010.html
[4] https://sites.google.com/site/singingvoiceseparationrpca/
[5] http://www.loria.fr/ aliutkus/kaml/
[6] http://bass-db.gforge.inria.fr/fasst/
[7] http://zafarrafii.com/repet.html
[8] http://www.audiolabs-erlangen.de/resources/2014-DAFx-Unison/

Table 1. Results for the D1 dataset: (a) The performance of T1 for the instantaneous mixtures averaged over datasets "test" and "test2" in 2 mics and the over dataset "test3" in 3 mics. (b) The performance of T2 for the convolutive mixtures averaged over "test" dataset in 2 mics and over "test3" dataset in 3 mics. SP and MU represents speech and music data, respectively.

(a)

System	2mic/3src (SP)			2mic/3src (MU)			2mic/4src (SP)			3mic/4src (SP)		
	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR
Sgouros [4]	7.6	18.8	8.6	8.3	18.4	9.4	5.6	15.6	6.5	6.6	19.1	7.0

(b)

System	2mic/3src (SP)				2mic/4src (SP)				3mic/4src (SP)			
	SDR	ISR	SIR	SAR	SDR	ISR	SIR	SAR	SDR	ISR	SIR	SAR
	OPS	TPS	IPS	APS	OPS	TPS	IPS	APS	OPS	TPS	IPS	APS
Bouafif [5]	-4.3	1.4	-1.9	8.6	-5.7	1.6	-3.6	8.2	–	–	–	–
	8.4	67.0	1.4	85.1	8.4	55.1	1.0	83.3	–	–	–	–
Nguyen	7.0	11.6	11.6	9.2	4.5	8.3	8.0	6.4	4.3	7.2	6.6	8.0
	40.9	65.3	55.9	58.0	36.9	62.2	51.0	48.7	35.6	62.2	53.3	47.0

Figures 2 and 3 show the box plots for the SDR, ISR, SIR, and SAR (in dB), for the vocals and the accompaniment, respectively, for the test subset. Outliers are not shown, median values are displayed, and higher values are better. As we can see, the separation performance is overall better for the accompaniment, as many songs feature weak vocals. Also, supervised systems typically achieved better results compared to unsupervised systems. Finally, depending on the systems, more or less large statistical dispersions are observed, meaning that different methods lead to different performances, depending on the songs, hence the need for a large-scale evaluation for music source separation.

Three methods were submitted to D4. Wang's method consists of an exhaustive search for estimating the sampling frequency mismatch and a state-of-the-art source separation technique [25]. Their results show the highest SIR but ISR is not so high. Miyabe's method consists of the maximum likelihood estimation of the sampling frequency mismatch [26] followed by auxiliary function based independent vector analysis [27]. Their results show the highest ISR. So, this combination would be interesting. Murase's system does not include the compensation of sampling frequency mismatch. It directly designs the time-frequency mask based on non-negative matrix factorization in the time-channel domain with sparse penalty added to [28]. It is robust to the sampling frequency mismatch, but the performance is limited due to using amplitude information only. Also, the results of realmix and simrefs are almost the same for all algorithms, which indicates that an effective evaluation was obtained by preparing the ground truth with time marking proposed in this task.

Table 2. Results for the D2 dataset (only for task T1)

systems	criteria	dev			test					
		Ca1	Sq1	Su1	Ca1	Ca2	Sq1	Sq2	Su1	Su2
López [7]	SDR	-	-	-	4.0	4.5	5.1	11.0	−3.8	3.9
	SIR	-	-	-	14.9	16.1	9.6	16.3	−1.6	8.8
	SAR	-	-	-	4.7	5.0	8.6	13.0	4.3	6.3
Ito [8]	SDR	7.2	8.9	4.9	8.1	7.8	10.8	13.8	6.7	7.6
	SIR	25.9	23.7	15.3	25.7	27.7	26.8	28.6	21.0	27.9
	SAR	7.2	9.2	5.6	8.2	7.8	11.0	14.0	6.9	7.7

Table 3. Results of T2 for the D4 dataset

Systems	criteria	3src			4src		
		realmix	sumrefs	mix	realmix	sumrefs	mix
Wang [25]	SDR	4.4	4.4	4.6	3.0	3.0	2.5
	ISR	4.8	4.9	5.2	3.5	3.6	3.3
	SIR	20.8	20.7	18.6	18.0	17.9	16.8
	SAR	12.8	12.9	13.9	11.0	11.2	10.9
Miyabe [26]	SDR	6.9	6.8	10.6	4.0	3.8	3.3
	ISR	11.2	11.1	15.1	8.8	8.5	7.3
	SIR	11.0	10.9	14.9	6.7	6.4	6.0
	SAR	11.7	11.6	15.5	7.8	7.6	7.4
Murase	SDR	2.7	2.6	2.4	0.9	0.8	1.0
	ISR	7.0	6.8	7.0	5.2	5.1	5.3
	SIR	5.2	4.6	4.2	1.7	1.6	2.3
	SAR	5.0	5.3	5.5	4.2	4.2	3.6

4 Conclusion

In this paper, we reported the tasks, datasets and evaluation criteria with the evaluation results in SiSEC 2015. Two new datasets were added in this SiSEC. We hope that these datasets and the evaluation results will be used in future research of the source separation field. Also, we have a plan to conduct web-based perceptual evaluation, which will be presented as follow-up report.

Acknowledgment We would like to thank Dr. Shigeki Miyabe for providing the new ASY dataset, and Mike Senior for giving us the permission to use the the MSD database for creating the MSD100 corpus.

References

1. N. Ono, Z. Koldovsky, S. Miyabe and N. Ito, "The 2013 Signal Separation Evaluation Campaign", in *Proc. MLSP*, Sept. 2013, pp. 1–6.
2. E. Vincent, R. Griboval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

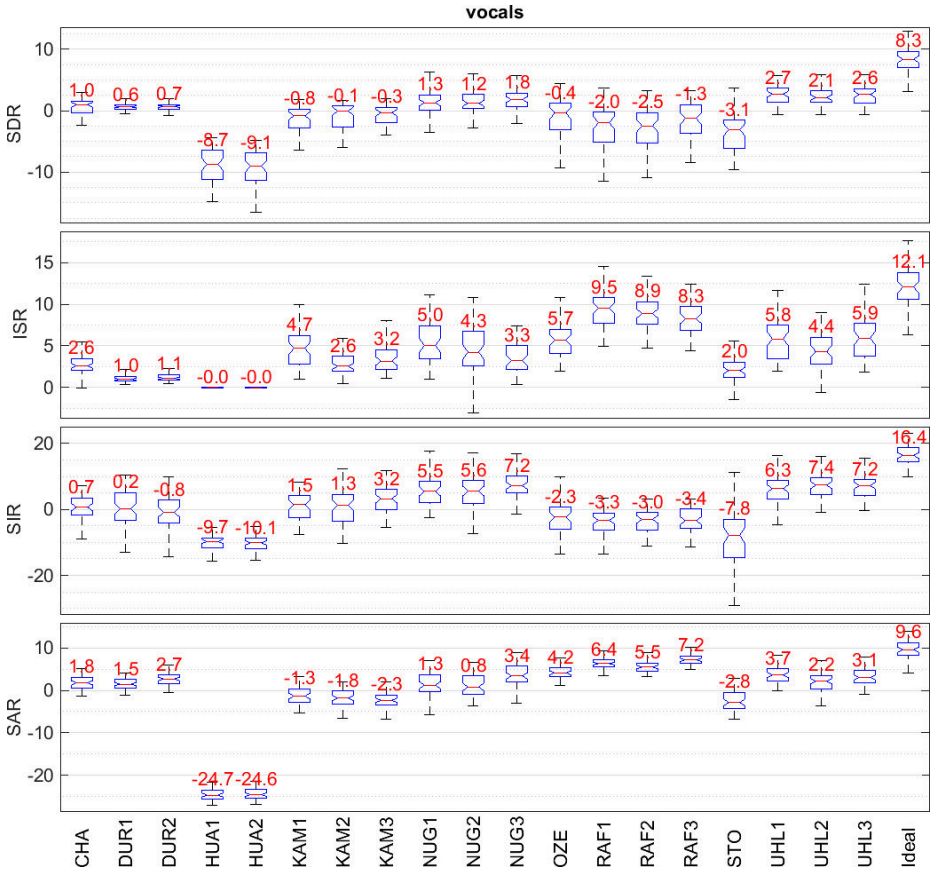

Fig. 2. Results of T2 for the D3 dataset (vocals).

3. V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Trans. ASLP*, vol. 19, no. 7, pp. 2046–2057, Sep. 2011.
4. N. Mitianoudis, "A Generalised Directional Laplacian Distribution: Estimation, Mixture Models and Audio Source Separation," *IEEE Trans. ASLP*, vol. 20, no. 9, pp. 2397–2408, 2012.
5. M. Bouafif and Z. Lachiri, "Multi-Sources Separation for Sound Source Localization," in *Proc. Interspeech* Sept. 2014, pp. 14–18.
6. H. Sawada, S. Araki and S. Makino, "Underdetermined Convolutive Blind Source Separation via Frequency Bin-Wise Clustering and Permutation Alignment," *IEEE Trans. ASLP*, vol. 19, no. 3, pp. 516–527, 2011.
7. A. R. López, N. Ono, U. Remes, K. Palomäki and M. Kurimo, "Designing Multichannel Source Separation Based on Single-Channel Source Separation," in *Proc. ICASSP*, Apr. 2015, pp. 469-473.
8. N. Ito, S. Araki, and T. Nakatani, "Permutation-free convolutive blind source separation via full-band clustering based on frequency-independent source presence priors," in *Proc. ICASSP*, May 2013, pp. 3238–3242.

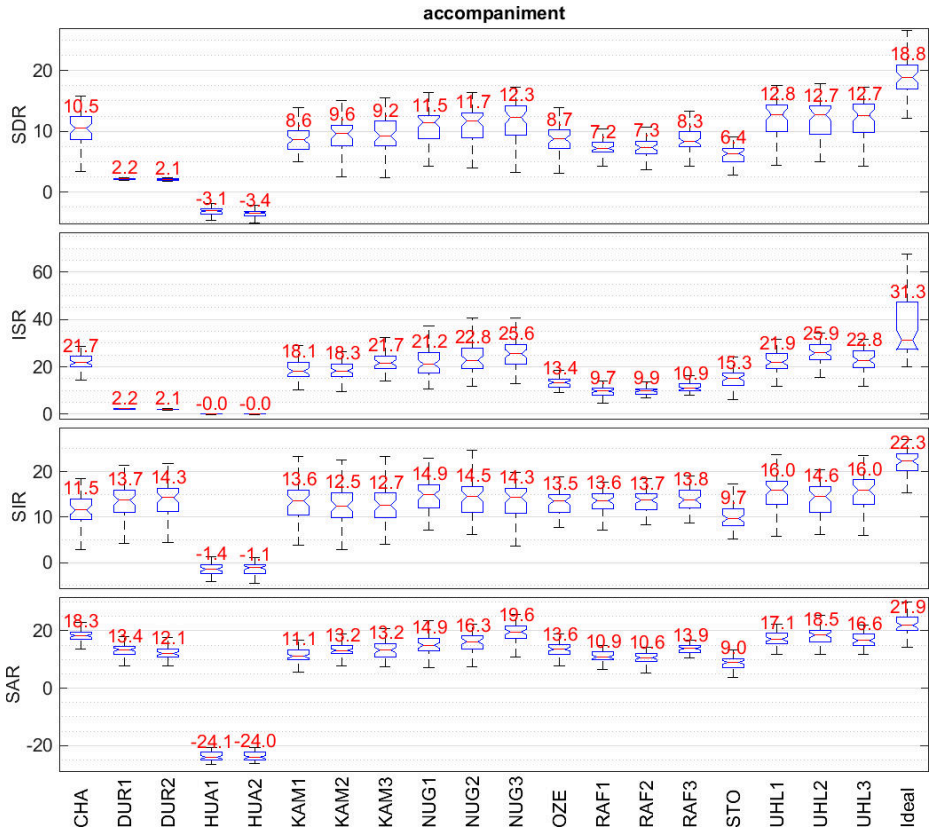

Fig. 3. Results of T2 for the D3 dataset (accompaniment).

9. Tak-Shing Chan, Tzu-Chun Yeh, Zhe-Cheng Fan, Hung-Wei Chen, Li Su, Yi-Hsuan Yang, and Roger Jang, "Vocal activity informed singing voice separation with the iKala dataset," in *Proc. ICASSP*, Apr. 2015, pp. 718–722.
10. Jean-Louis Durrieu, Bertrand David, and Gaël Richard, "A musically motivated mid-level representation for pitch estimation and musical audio source separation," *IEEE Journal on Selected Topics on Signal Processing*, vol. 5, no. 6, pp. 1180–1191, Oct. 2011.
11. Po-Sen Huang, Scott Deeann Chen, Paris Smaragdis, and Mark Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *Proc. ICASSP*, Mar. 2012, pp. 57–60.
12. Antoine Liutkus, Derry FitzGerald, Zafar Rafii, Bryan Pardo, and Laurent Daudet, "Kernel additive models for source separation," *IEEE Trans. SP*, vol. 62, no. 16, pp. 4298–4310, August 2014.
13. Antoine Liutkus, Derry FitzGerald, Zafar Rafii, and Laurent Daudet, "Scalable audio separation with light kernel additive modelling," in *Proc. ICASSP*, Apr. 2015, pp. 76–80.

14. Aditya A. Nugraha, Antoine Liutkus, and Emmanuel Vincent, "Multichannel audio source separation with deep neural networks," Research Report RR-8740, Inria, 2015.
15. Alexey Ozerov, Emmanuel Vincent, and Frédéric Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Trans. ASLP*, vol. 20, no. 4, pp. 1118–1133, Oct. 2012.
16. Yann Salaün, Emmanuel Vincent, Nancy Bertin, Nathan Souviraà-Labastie, Xabier Jaureguiberry, Dung T. Tran, and Frédéric Bimbot, "The flexible audio source separation toolbox version 2.0," in *Proc. ICASSP*, May 4-9 2014.
17. Zafar Rafii and Bryan Pardo, "REpeating Pattern Extraction Technique (REPET): A simple method for music/voice separation," *IEEE Trans. ASLP*, vol. 21, no. 1, pp. 71–82, January 2013.
18. Antoine Liutkus, Zafar Rafii, Roland Badeau, Bryan Pardo, and Gaël Richard, "Adaptive filtering for music/voice separation exploiting the repeating musical structure," in *Proc. ICASSP*, Mar. 2012, pp. 53–56.
19. Zafar Rafii and Bryan Pardo, "Music/voice separation using the similarity matrix," in *Proc. ISMIR*, Oct. 2012, pp. 583–588.
20. Zafar Rafii, Antoine Liutkus, and Bryan Pardo, "REPET for background/foreground separation in audio," in *Blind Source Separation*, Ganesh R. Naik and Wenwu Wang, Eds., Signals and Communication Technology, chapter 14, pp. 395–411. Springer Berlin Heidelberg, 2014.
21. Justin Salamon and Emilia Gómez, "Melody extraction from polyphonic music signals using pitch contour characteristics," in *IEEE Trans. ASLP*, vol. 20, no. 6, pp. 1759–1770, Aug. 2012.
22. Fabian-Robert Stöter, Stefan Bayer, and Bernd Edler, "Unison Source Separation," in *Proc. DAFx*, Sep. 2014.
23. Stefan Uhlich, Franck Giron, and Yuki Mitsufuji, "Deep neural network based instrument extraction from music," in *Proc. ICASSP*, Apr. 2015, pp. 2135–2139.
24. Hakan Erdogan, John R. Hershey, Shinji Watanabe, and Jonathan Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. ICASSP*, Apr. 2015, pp. 708–712.
25. L. Wang, "Multi-band multi-centroid clustering based permutation alignment for frequency-domain blind speech separation," *Digit. Signal Process.*, vol. 31, pp. 79–92, Aug. 2014.
26. S. Miyabe, N. Ono and S. Makino, "Blind compensation of interchannel sampling frequency mismatch for ad hoc microphone array based on maximum likelihood estimation," *Elsevier Signal Processing*, vol. 107, pp. 185–196, Feb. 2015.
27. N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. WASPAA*, Oct. 2011, pp. 189–192.
28. H. Chiba, N. Ono, S. Miyabe, Y. Takahashi, T. Yamada and S. Makino, "Amplitude-based speech enhancement with nonnegative matrix factorization for asynchronous distributed recording," in *Proc. IWAENC*, Sept. 2014, pp. 204–208.