Datasets

- 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

Tasks

- T1 Single-channel source estimation
- T2 Multichannel source image estimation

One algorithm was submitted to D1 with T2. The results are described in Tables 1–4 comparing with those in past SiSECs. Wood's algorithm [1] utilizes both generalized cross correlation (GCC) [10] and nonnegative matrix factorization (NMF) [11]. GCC was previously used for sound source localization in reverberant environments [12]. NMF is a famous mathematical framework for many applications, especially in the source separation task. For the acoustic signals, NMF can extract some spectral patterns (bases) and their activations (time-varying gains), and the source separation based on NMF is achieved by clustering the bases into each source. Wood et al. combined GCC with NMF to localize individual bases over time, such that they may be attributed to individual sources. Each source is then reconstructed independently using only the associated bases. More precisely, the NMF decomposition is first performed on a magnitude spectrogram of the mixture signal with channels concatenated in time. Then, each basis is subsequently attributed to a single source at each time according to its spatial origin estimated by GCC.

The computational times of Wood's algorithm were between 6 and 7 minutes per mixture, where they used a dual 2.8 GHz Intel Xeon E5462 quad-core processor with 16GB of RAM. From the comparison of the results, Wood's algorithm could not outperform the best ever performance. For the mixture "test_female4_liverec_250ms_5cm_mix.wav" (the second column in Table 4) the separation was failed because GCC localization could not find all the source locations due to the reverberation.

1. REFERENCES

[1] S. Wood and J. Rouat, "Blind speech separation with GCC-NMF," in *Proc. Interspeech*, 2016 (to appear).

[2] M. Bouafif and Z. Lachiri, "Multi-sources separation for sound source localization," in *Proc. Interspeech* pp. 14–18, 2014.

[3] J. Cho and C. D. Yoo, "Underdetermined convolutive BSS: Bayes risk minimization based on a mixture of super-Gaussian posterior approximation," in *IEEE Trans. Audio, Speech, and Language Processing*, vol. 23, no. 5, pp. 828–839, 2011.

[4] K. Adiloglu and E. Vincent, "Variational Bayesian inference for source separation and robust feature extraction," *Technical Report*, INRIA, https://hal.inria.fr/hal-00726146, 2012

[5] Y. Hirasawa, N. Yasuraoka, T. Takahashi, T. Ogata and H. Okuno, "A GMM sound source model for blind speech separation in under-determined conditions," in *Proc. International Conference on Latent Variable Analysis and Signal Separation*, 2012.

[6] K. Iso, S. Araki, S. Makino, T. Nakatani, H. Sawada, T. Yamada and A. Nakamura, "Blind source separation of mixed speech in a high reverberation environment," in *Proc. Hands-free Speech Communication and Microphone Arrays*, pp. 36–39, 2011.

[7] J. Cho, J. Choi and C. D. Yoo, "Underdetermined convolutive blind source separation using a novel mixing matrix estimation and MMSE-based source estimation," in *Proc. IEEE International Workshop on Machine Learning for Signal Processing*, 2011.

[8] F. Nesta and M. Omologo "Convolutive underdetermined source separation through weighted Interleaved ICA and spatio-temporal correlation," in *Proc. International Conference on Latent Variable Analysis and Signal Separation*, 2012

[9] A. Ozerov, E. Vincent and F. Bimbot "A general flexible framework for the handling of prior information in audio source separation," in *IEEE Trans. Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1118–1133, 2011.

[10] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," in *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[11] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," in *Nature*, vol. 401, pp. 788–791, 1999.

[12] C. Blandin, A. Ozerov and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," in *Signal Processing*, vol. 92, no. 8, pp. 1950–1960, 2012.

Table 1. Results for database D1 and task T2 for the convolutive mixtures averaged over sources: live-recorded data with 1 m microphone spacing and 130 ms reverberation time in dataset "test"

System	2mic/3src (female speech)				2mic/4src (female speech)				2mic/3src (male speech)				2mic/4src (male speech)			
	SDR	ISR	SIR	SAR	SDR	ISR	SIR	SAR	SDR	ISR	SIR	SAR	SDR	ISR	SIR	SAR
	OPS	TPS	IPS	APS	OPS	TPS	IPS	APS	OPS	TPS	IPS	APS	OPS	TPS	IPS	APS
Wood [1]	3.3	7.4	4.3	7.2	2.3	5.4	2.9	5.1	3.6	7.1	5.1	6.5	3.1	6.0	4.8	5.3
(SiSEC 2016)	8.8	6.6	6.9	17.0	15.8	17.6	19.8	26.0	9.1	7.4	8.6	20.2	13.9	17.0	20.8	28.4
Bouafif [2]	-4.3	1.5	-2.3	9.7	-5.8	1.1	-4.1	10.3	-4.4	1.4	-1.5	7.5	-5.6	2.1	-3.2	6.0
(SiSEC 2015)	8.4	72.0	1.5	85.7	8.4	62.3	1.2	84.4	8.4	61.9	1.4	84.4	8.4	47.9	0.8	82.3
Nguyen	7.3	11.6	11.6	10.0	5.2	9.1	9.1	7.1	6.7	11.6	11.7	8.4	3.8	7.5	6.8	5.7
(SiSEC 2015)	40.0	62.3	53.7	62.8	38.9	65.0	52.9	50.9	41.9	68.4	58.0	53.2	34.8	59.4	49.1	46.5
Cho [3]	6.2	10.5	9.0	9.7	4.3	8.0	6.6	7.3	6.5	11.0	10.1	9.6	4.5	8.2	6.8	6.9
(SiSEC 2013)	36.4	63.0	49.5	64.0	36.9	64.5	43.1	56.1	28.6	60.9	44.0	70.8	34.8	60.6	39.1	55.5
Adiloglu [4]	2.4	7.4	3.5	8.2	2.1	6.0	2.9	5.9	3.8	8.8	7.2	8.2	3.3	7.1	5.6	6.3
(SiSEC 2013)	20.5	50.9	37.6	70.3	32.9	49.5	36.6	56.8	26.6	56.1	43.4	71.1	36.2	57.5	42.6	56.2
Hirasawa [5]	2.3	4.5	3.4	4.7	1.6	3.7	2.5	3.3	2.0	4.2	3.0	3.9	1.7	3.7	2.5	2.7
(SiSEC 2011)	26.7	38.6	45.8	41.2	21.8	25.9	41.9	32.9	26.5	45.1	44.7	44.4	23.3	33.4	42.5	38.1
Iso [6]	7.3	11.2	11.5	11.0	–	–	–	–	1.9	5.5	2.5	7.0	–	–	–	–
(SiSEC 2011)	24.8	56.7	46.3	73.8	–	–	–	–	18.5	28.6	16.6	46.0	–	–	–	–
Cho [7]	3.8	8.3	5.4	8.1	1.9	5.8	1.9	6.1	4.4	9.2	7.8	7.5	2.4	5.9	3.7	5.4
(SiSEC 2011)	18.2	29.2	18.9	51.0	30.5	48.8	31.5	55.2	26.9	56.2	35.4	67.6	28.5	37.1	25.7	51.8
Nesta (1) [8]	6.3	8.8	10.3	10.3	2.8	5.3	4.3	6.6	6.2	8.8	10.6	9.2	4.1	6.8	7.2	6.4
(SiSEC 2011)	27.4	56.8	49.7	71.7	35.9	59.0	43.7	56.5	32.7	60.9	48.9	67.5	36.3	60.5	46.3	54.2
Nesta (2) [8]	7.7	12.1	12.3	10.7	3.0	6.8	5.3	6.2	7.1	12.0	12.3	9.2	4.7	8.6	8.4	6.4
(SiSEC 2011)	38.3	59.6	54.7	62.8	36.2	58.5	48.9	52.1	42.7	65.4	59.0	55.5	36.4	60.4	54.2	45.9
Ozerov [9]	3.2	8.5	4.6	7.4	2.6	6.3	4.4	6.1	2.8	6.7	5.1	7.2	2.8	6.8	5.0	5.8
(SiSEC 2011)	26.0	51.8	40.8	64.9	26.8	32.5	28.7	36.3	26.1	53.4	32.2	66.2	37.7	53.6	47.1	51.3

Table 2. Results for database D1 and task T2 for the convolutive mixtures averaged over sources: live-recorded data with 5 cm microphone spacing and 130 ms reverberation time in dataset "test"

System	2mic/3src (female speech)				2mic/4src (female speech)				2mic/3src (male speech)				2mic/4src (male speech)			
	SDR	ISR	SIR	SAR	SDR	ISR	SIR	SAR	SDR	ISR	SIR	SAR	SDR	ISR	SIR	SAR
	OPS	TPS	IPS	APS	OPS	TPS	IPS	APS	OPS	TPS	IPS	APS	OPS	TPS	IPS	APS
Wood [1]	3.3	8.1	8.3	8.9	3.1	6.5	7.4	6.9	2.8	6.9	7.3	6.7	2.9	6.0	7.2	5.7
(SiSEC 2016)	8.6	5.3	7.6	17.7	9.5	6.3	10.7	22.0	9.2	5.7	12.7	23.6	11.1	8.7	18.9	25.9
Bouafif [2]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
(SiSEC 2015)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Nguyen	6.8	12.4	11.2	10.4	4.2	7.9	7.6	6.4	6.3	10.7	10.9	8.5	4.4	8.2	8.1	6.2
(SiSEC 2015)	37.8	61.9	57.5	66.4	37.5	60.2	48.1	54.1	21.6	25.4	24.4	36.0	34.6	58.9	50.3	46.7
Cho [3]	8.7	13.1	12.8	12.2	4.2	7.5	6.3	7.6	6.5	10.7	10.3	9.8	4.6	8.0	6.5	7.8
(SiSEC 2013)	18.6	54.3	51.1	80.7	31.3	58.2	36.8	63.4	21.7	57.7	48.2	78.7	30.7	56.4	33.6	61.6
Adiloglu [4]	3.5	7.5	5.7	9.0	2.6	7.0	6.2	7.3	3.7	8.4	8.3	7.4	2.7	6.4	5.5	5.9
(SiSEC 2013)	22.9	51.4	45.7	70.2	37.2	56.4	43.6	56.9	33.4	61.1	48.7	67.3	36.2	53.8	42.8	54.3
Hirasawa [5]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
(SiSEC 2011)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Iso [6]	8.4	12.9	12.2	12.3	–	–	–	–	5.8	9.8	8.8	9.4	–	–	–	–
(SiSEC 2011)	23.3	54.1	56.1	77.4	–	–	–	–	22.0	41.2	29.2	62.9	–	–	–	–
Cho [7]	5.2	9.7	9.0	9.3	2.2	5.3	2.9	6.8	5.6	11.1	9.6	9.1	2.3	6.1	3.6	5.8
(SiSEC 2011)	15.5	46.1	37.9	73.2	30.8	53.0	35.8	61.2	20.4	58.2	43.9	78.1	32.8	55.0	37.4	58.0
Nesta (1) [8]	6.4	10.5	9.9	10.9	4.8	8.0	8.9	7.7	6.4	9.8	10.9	9.3	3.9	6.9	7.0	6.6
(SiSEC 2011)	30.9	58.5	59.6	74.5	38.1	62.3	49.7	57.4	36.0	62.5	57.6	68.6	36.3	61.4	49.8	52.7
Nesta (2) [8]	6.6	13.2	11.6	10.3	4.9	8.9	9.8	7.4	6.8	11.7	12.5	8.5	3.9	7.7	8.2	6.2
(SiSEC 2011)	47.3	63.3	64.8	63.6	40.5	61.0	57.2	50.9	47.8	69.0	66.8	49.3	36.6	60.2	56.9	45.6
Ozerov [9]	4.7	9.1	7.4	8.8	2.9	7.5	6.2	7.2	4.6	9.3	9.4	8.0	2.5	6.3	5.2	5.8
(SiSEC 2011)	28.7	55.9	51.1	69.4	39.0	56.9	47.9	54.9	32.9	60.8	50.3	68.3	36.1	56.0	48.3	50.7

Table 3. Results for database D1 and task T2 for the convolutive mixtures averaged over sources: live-recorded data with 1 m microphone spacing and 250 ms reverberation time in dataset "test"

System	2mic/3src (female speech)				2mic/4src (female speech)				2mic/3src (male speech)				2mic/4src (male speech)			
	SDR	ISR	SIR	SAR	SDR	ISR	SIR	SAR	SDR	ISR	SIR	SAR	SDR	ISR	SIR	SAR
	OPS	TPS	IPS	APS	OPS	TPS	IPS	APS	OPS	TPS	IPS	APS	OPS	TPS	IPS	APS
Wood [1]	3.2	6.7	4.7	6.8	2.2	5.0	2.8	4.8	3.1	6.5	4.3	6.6	2.5	5.2	3.1	4.8
(SiSEC 2016)	10.6	8.6	9.0	23.3	27.4	43.7	35.3	47.1	9.7	8.8	9.9	24.2	29.6	47.9	41.7	44.5
Bouafif [2]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
(SiSEC 2015)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Nguyen	6.1	9.9	9.3	9.6	4.0	7.5	7.1	7.1	5.9	10.1	9.8	8.2	2.5	5.8	4.1	5.4
(SiSEC 2015)	37.1	63.0	48.2	59.0	34.7	60.3	47.6	49.9	40.0	65.8	53.1	53.7	31.8	50.8	43.1	48.0
Cho [3]	5.5	9.5	8.1	9.4	4.3	7.8	6.8	7.5	5.5	9.5	8.2	9.1	3.2	6.6	4.7	6.2
(SiSEC 2013)	35.6	62.9	43.4	59.0	33.3	59.0	38.3	52.3	36.0	61.5	44.8	58.7	35.1	57.0	42.8	50.8
Adiloglu [4]	3.0	7.0	5.5	8.1	0.7	4.3	0.9	4.8	3.4	7.1	5.8	8.4	1.5	5.0	2.1	5.2
(SiSEC 2013)	28.4	53.7	35.2	60.8	29.2	46.4	29.4	53.3	26.4	51.4	31.8	63.0	32.7	52.2	36.1	56.1
Hirasawa [5]	2.2	4.2	4.3	4.0	1.2	3.2	0.9	2.6	1.7	3.8	2.8	3.6	0.9	3.0	0.4	1.9
(SiSEC 2011)	22.6	32.6	46.8	38.1	19.5	23.6	41.6	32.8	24.6	36.1	44.0	41.2	20.2	26.3	41.6	34.5
Iso [6]	6.1	9.8	8.7	10.9	–	–	–	–	5.5	9.4	8.5	9.1	–	–	–	–
(SiSEC 2011)	30.4	59.6	45.1	64.8	–	–	–	–	30.9	54.5	35.0	59.8	–	–	–	–
Cho [7]	3.2	7.4	4.4	8.1	0.0	3.1	-0.7	5.8	4.2	8.8	6.7	8.0	0.9	4.2	1.2	5.2
(SiSEC 2011)	22.0	27.8	20.8	43.6	21.7	24.7	20.0	40.5	37.4	63.3	46.4	55.5	25.2	32.4	25.0	46.4
Nesta (1) [8]	4.3	6.5	7.9	8.4	2.8	5.2	5.3	6.2	4.9	7.5	9.1	7.5	3.5	5.9	6.6	5.1
(SiSEC 2011)	38.1	63.1	52.0	56.3	35.5	54.7	49.5	45.8	41.2	63.5	55.0	52.5	35.7	56.3	53.6	42.2
Nesta (2) [8]	6.0	10.2	10.4	10.2	3.4	6.9	6.3	7.2	6.2	10.3	10.4	8.6	4.7	8.3	8.3	6.3
(SiSEC 2011)	37.3	60.8	50.5	60.2	33.6	49.5	45.0	50.1	39.8	60.1	52.1	55.2	35.7	54.5	51.1	49.6
Ozerov [9]	3.6	8.2	7.4	7.4	1.5	5.1	2.5	4.7	6.0	10.4	9.9	8.8	2.2	5.9	3.8	5.4
(SiSEC 2011)	36.0	63.5	48.1	56.2	30.6	47.5	38.1	49.5	39.6	61.3	51.7	58.2	37.4	55.9	50.3	51.7

Table 4. Results for database D1 and task T2 for the convolutive mixtures averaged over sources: live-recorded data with 5 cm microphone spacing and 250 ms reverberation time in dataset "test"

System	2mic/3src (female speech)				2mic/4src (female speech)				2mic/3src (male speech)				2mic/4src (male speech)			
	SDR	ISR	SIR	SAR	SDR	ISR	SIR	SAR	SDR	ISR	SIR	SAR	SDR	ISR	SIR	SAR
	OPS	TPS	IPS	APS	OPS	TPS	IPS	APS	OPS	TPS	IPS	APS	OPS	TPS	IPS	APS
Wood [1]	3.6	8.2	7.0	7.4	–	–	–	–	3.7	7.5	7.0	6.8	1.6	4.9	4.5	5.1
(SiSEC 2016)	34.0	55.8	45.3	53.8	–	–	–	–	35.4	57.5	49.2	49.5	11.8	13.6	20.5	29.6
Bouafif [2]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
(SiSEC 2015)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Nguyen	5.8	10.2	9.7	9.3	2.9	6.4	4.8	6.3	5.6	9.9	8.8	7.9	3.5	6.8	5.6	5.7
(SiSEC 2015)	36.8	63.3	49.7	60.4	35.4	57.9	47.9	51.7	42.1	65.2	55.1	55.1	34.3	59.1	47.3	48.5
Cho [3]	6.3	10.2	9.5	10.2	4.6	7.7	7.1	7.6	6.1	10.3	8.9	9.4	3.8	6.8	5.0	7.1
(SiSEC 2013)	25.5	57.9	40.6	71.7	32.2	60.2	36.5	57.3	27.1	58.1	42.2	68.7	34.7	61.2	39.2	55.5
Adiloglu [4]	3.6	8.3	6.7	8.1	0.7	4.8	-0.3	5.6	4.5	8.8	8.2	8.7	2.0	5.8	3.4	5.7
(SiSEC 2013)	35.7	60.3	47.5	62.7	17.7	20.1	17.9	40.6	31.8	59.5	42.6	63.5	33.4	53.3	39.4	55.8
Hirasawa [5]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
(SiSEC 2011)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Iso [6]	6.6	10.6	10.2	10.7	–	–	–	–	5.6	10.1	8.2	9.8	–	–	–	–
(SiSEC 2011)	24.5	56.1	40.2	70.8	–	–	–	–	26.0	54.3	44.4	70.6	–	–	–	–
Cho [7]	5.3	9.4	8.6	9.4	1.2	4.4	1.7	5.6	4.6	9.9	7.2	9.1	2.4	5.8	3.6	5.8
(SiSEC 2011)	27.0	58.4	40.5	70.0	31.9	49.6	33.7	54.3	26.5	56.9	46.0	68.6	30.7	50.1	32.7	55.0
Nesta (1) [8]	6.6	10.4	11.1	9.8	2.3	5.6	3.9	5.6	5.4	9.2	8.8	8.0	2.9	6.2	4.9	5.2
(SiSEC 2011)	39.1	65.4	53.1	61.3	37.0	57.9	48.7	48.1	37.0	61.2	51.3	61.2	36.9	58.9	50.7	46.9
Nesta (2) [8]	7.8	12.3	13.4	10.9	2.6	6.3	4.9	6.2	6.0	10.1	9.9	8.3	3.5	7.1	6.2	5.8
(SiSEC 2011)	34.4	61.6	49.7	63.0	34.4	51.9	44.5	51.3	39.1	59.8	52.5	58.1	36.5	54.8	48.7	51.3
Ozerov [9]	4.0	9.0	8.2	8.0	2.5	6.9	4.8	5.6	4.7	9.1	8.9	8.8	2.3	6.2	4.2	5.4
(SiSEC 2011)	35.5	60.6	48.6	64.4	27.6	41.8	31.3	50.5	29.1	57.6	44.6	65.6	35.8	57.0	43.6	54.2