SIMILARITY SEARCH-BASED BLIND SOURCE SEPARATION

Hiroshi Sawada and Kazuo Aoyama

NTT Communication Science Laboratories, NTT Corporation

ABSTRACT

In this paper, we propose a new method for blind source separation, where we perform similarity search for a prepared clean speech database. The purpose of this mechanism is to separate short utterances that we frequently encounter in a real-world situation. The new method employs a local Gaussian model (LGM) for the probability density functions of separated signals, and updates the LGM variance parameters by using the similarity search results. Experimental results show that the method performed very well in an ideal situation where we employ a close database that contains the source components used for the mixtures. In more realistic situations where an open database was used, the separation performance degraded to a certain degree, but was still better than existing methods.

Index Terms— frequency domain blind source separation, local Gaussian model, variance parameter, similarity search, Itakura-Saito divergence

1. INTRODUCTION

Blind source separation (BSS) is a technique for estimating individual source components from their mixtures [1, 2]. When the sources are sounds in a real room environment, they are mixed with delay and reverberations, (i.e., convolutive). A typical approach to such convolutive mixtures is frequency domain BSS [3] where we transform the time-domain mixtures into time-frequency domain complex coefficients by using short-time Fourier transform (STFT).

For solving the frequency domain BSS problem, most existing methods rely on statistical properties of sources, such as independence or low-rank structures. Regarding independence, independent component analysis (ICA) [4] could be employed to separate the mixtures in each frequency bin. However, we need to solve the permutation problem [5] afterwards to align the order ambiguities of ICA solutions. Independent vector analysis (IVA) [6–10] separates the mixtures so that the permutation problem is solved automatically by considering the independence of vectors that span all frequency bins. Regarding low-rank structures, independent low-rank matrix analysis (ILRMA) [11] assumes that a separated signal has a spectrogram structure of a lower rank than a mixture has. All such methods need to obtain enough amount of observations, say sound mixtures of 3 seconds or more, so that the statistical properties work effectively.

This paper proposes a new BSS method that separates shorter observations, say sound mixtures of 2 seconds or less, by performing similarity search on a prepared clean source database. We human can separate sound mixtures easily if there is something familiar to us in the mixtures. The new method is inspired by and tries to imitate such human ability.

Preparing a source database is similar to the situations of supervised learning. For the task of source separation, many supervised methods have been proposed such as basis decomposition using nonnegative matrix factorization (NMF) [12], and more recent deep neural network (DNN) based methods [13–19]. These methods need a supervised training phase that sometimes is very time consuming. The proposed method does not need such a time consuming training phase.

The proposed method, as well as some of the aforementioned existing methods, employs a local Gaussian model (LGM) [20–22] for the probability density functions of separated signals. The next section formulates the problem of frequency domain BSS, and reviews the existing methods employing LGMs. Section 3 explains the proposed method as a natural extension of the LGM-based BSS methods for utilizing a clean source database. Section 4 reports experimental results. Section 5 concludes this paper.

2. FREQUENCY DOMAIN BSS

Let $x_{ftm} \in \mathbb{C}$ be the STFT coefficient of the mixture at frequency bin f, time frame t, and sensor m. Having M sensors, these coefficients are summarized in a vector form as $\mathbf{x}_{ft} = [x_{ft1}, \ldots, x_{ftM}]^\mathsf{T} \in \mathbb{C}^M$. Having F frequency bins and T time frames, the purpose of frequency domain BSS is to separate the mixtures $\mathbf{x}_{ft}$, $f = 1, \ldots, F$, $t = 1, \ldots, T$ by

$$\mathbf{y}_{ft} = \mathbf{W}_f \mathbf{x}_{ft} \tag{1}$$

where $\mathbf{y}_{ft} = [y_{ft1}, \ldots, y_{ftN}]^\mathsf{T} \in \mathbb{C}^N$ is the vector of N separated signals and $\mathbf{W}_f$ is the $N \times M$ frequency-bin specific separation matrix

$$\mathbf{W}_f = \begin{bmatrix} \mathbf{w}_{f1}^\mathsf{H} \\ \vdots \\ \mathbf{w}_{fN}^\mathsf{H} \end{bmatrix} \tag{2}$$

with $\mathbf{w}_{fn}$ being an M dimensional complex-valued vector.

2.1. IVA with local Gaussian model (LGM)

IVA optimizes the separation matrices $\{\mathbf{W}_f\}_{f=1}^F$ by minimizing the objective function

$$\mathcal{J}(\{\mathbf{W}_f\}_{f=1}^F) = \sum_{t=1}^{T}\sum_{n=1}^{N} G(\tilde{\mathbf{y}}_{tn}) - 2T\sum_{f=1}^{F}\log|\det \mathbf{W}_f|\,. \tag{3}$$

The factor 2 of the second term comes from the complex-valued density transformation $p(\mathbf{x}_{ft}|\mathbf{W}_f) = |\det \mathbf{W}_f|^2 p(\mathbf{y}_{ft})$ associated with (1). $G(\tilde{\mathbf{y}}_{tn})$ of the first term is a contrast function defined for a vector $\tilde{\mathbf{y}}_{tn} = [y_{1tn}, \ldots, y_{Ftn}]^\mathsf{T}$ that spans all the frequency bins $f = 1, \ldots, F$. Generally G is defined as $G(\tilde{\mathbf{y}}) = -\log p(\tilde{\mathbf{y}})$ with the probability density function $p(\tilde{\mathbf{y}})$. Considering the connection to the proposed method, we introduce an LGM

$$p(\tilde{\mathbf{y}}_{tn}) = \prod_{f=1}^{F}\mathcal{N}(y_{ftn}|0, v_{tn}) = \prod_{f=1}^{F}\frac{1}{\pi v_{tn}}\exp\left(-\frac{|y_{ftn}|^2}{v_{tn}}\right) \tag{4}$$

where v_{tn} is the variance parameter shared among all frequency bins. Then, the contrast function becomes

$$G(\tilde{\mathbf{y}}_{tn}) = \sum_{f=1}^{F}\left(\frac{|y_{ftn}|^2}{v_{tn}} + \log v_{tn}\right) \tag{5}$$

by omitting the constant term $F\log\pi$.

The objective function (3) with (5)

$$\mathcal{J}(\{\mathbf{W}_f\}_{f=1}^F, \{\{v_{tn}\}_{t=1}^T\}_{n=1}^N) \tag{6}$$

can be minimized by alternatively updating $\{\mathbf{W}_f\}_{f=1}^F$ and $\{\{v_{tn}\}_{t=1}^T\}_{n=1}^N$.

The variance parameters are updated by

$$v_{tn} \leftarrow \frac{1}{F}\sum_{f=1}^{F}|y_{ftn}|^2 \tag{7}$$

as a solution of the partial derivative of $\mathcal{J}$ with respect to v_{tn} being zero.

The separation matrices $\{\mathbf{W}_f\}_{f=1}^F$ are optimized in a frequency f bin-wise manner. Although a typical approach is to use natural gradient [23], we employ more recent ideas [8, 9, 24]. At first, we calculate a source n specific weighted covariance matrix

$$\mathbf{U}_{fn} = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{v_{tn}}\mathbf{x}_{ft}\mathbf{x}_{ft}^\mathsf{H} \tag{8}$$

using the updated variance parameters v_{tn}. Then, we solve $N \times N$ simultaneous equations

$$\mathbf{w}_{fk}^\mathsf{H}\mathbf{U}_{fn}\mathbf{w}_{fn} = \delta_{kn} \tag{9}$$

with δ_{kn} being the Kronecker delta for $k, n = 1, \ldots, N$. This problem has been formulated as hybrid exact-approximate diagonalization (HEAD) [24] for $\mathbf{U}_{f1}, \ldots, \mathbf{U}_{fN}$. An efficient way [8] to solve HEAD is to calculate

$$\mathbf{w}_{fn} \leftarrow (\mathbf{W}_f\mathbf{U}_{fn})^{-1}\mathbf{e}_n \tag{10}$$

for each n, where $\mathbf{e}_n$ is a vector whose nth element is one and all the others are zero, and update as

$$\mathbf{w}_{fn} \leftarrow \frac{\mathbf{w}_{fn}}{\sqrt{\mathbf{w}_{fn}^\mathsf{H}\mathbf{U}_{fn}\mathbf{w}_{fn}}} \tag{11}$$

to accommodate a HEAD constraint $\mathbf{w}_{fn}^\mathsf{H}\mathbf{U}_{fn}\mathbf{w}_{fn} = 1$. Since $\mathbf{W}_f$ consists of $\mathbf{w}_{f1}, \ldots, \mathbf{w}_{fN}$ as (2), the updates (10) and (11) should be iterated a few times to converge.

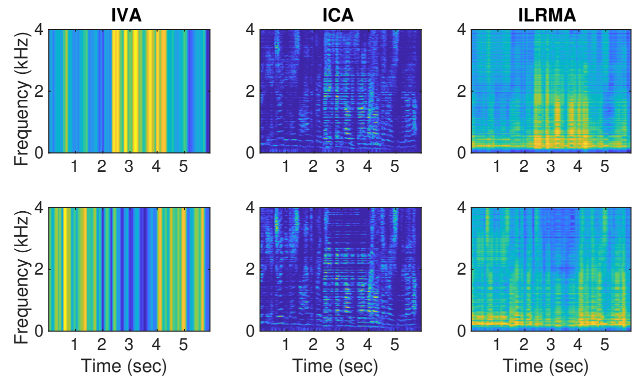

Fig. 1. Estimated variance parameter examples (log scale, large values colored in yellow) for 6-second mixtures. Each of two rows corresponds to each of two separated signals.

2.2. Examples of variance parameter estimations

Figure 1 shows some estimated variance parameters. IVA estimates time varying source activities with v_{tn} but not frequency structures. To estimate time-frequency structures, we can employ ICA with LGM where the probability density function $p(\tilde{\mathbf{y}})$ is expressed with frequency specific variances v_{ftn} as

$$p(\tilde{\mathbf{y}}_{tn}) = \prod_{f=1}^{F}p(y_{ftn}) = \prod_{f=1}^{F}\mathcal{N}(y_{ftn}|0, v_{ftn})\,, \tag{12}$$

$$\mathcal{N}(y_{ftn}|0, v_{ftn}) = \frac{1}{\pi v_{ftn}}\exp\left(-\frac{|y_{ftn}|^2}{v_{ftn}}\right). \tag{13}$$

However, permutation problems occur as shown in the second column of Fig. 1. This is because the joint density (4) is decomposed into the densities (12) of each frequency bin. The third column shows the variances v_{ftn} obtained by ILRMA which employs a low-rank model

$$v_{ftn} = \textstyle\sum_{k=1}^{K} b_{fnk}a_{tnk}\,, \quad b_{fnk}, a_{tnk} \geq 0 \tag{14}$$

where $[b_{1nk}, \ldots, b_{Fnk}]^\mathsf{T}$ represents the kth basis of the nth separation and a_{tnk} represents its activation for each time frame t. With 6-second mixtures, the low-rank structure ($K = 5$ in this case) is moderately well estimated.

3. SIMILARITY SEARCH-BASED METHOD

For the frequency domain BSS to work properly, the variance parameters v_{ftn} need to be constrained to solve the permutation problem. At the same time, v_{ftn} should have some flexibility to represent rich frequency structures of sources. ILRMA is suitable for these purposes as long as the observation length is long enough (e.g., 6 seconds). The main idea of the proposed method is to utilize a prepared clean source database $\mathbb{S}$ for estimating the variance parameters v_{ftn} well even if the observation length is short (e.g., 2 seconds).

3.1. Preparing clean source database

Let us have L clean source signals, $l = 1, \ldots, L$, for constructing the database $\mathbb{S}$. By applying STFT, we have coefficients $s_{ftl} \in \mathbb{C}$ in the time-frequency domain with $t = 1, \ldots, T_l$ being source l specific time frames. Let us define a vector that contains the power spectra of all frequency bins

$$\breve{\mathbf{s}} = [|s_1|^2, \ldots, |s_F|^2]^\mathsf{T} . \tag{15}$$

Now we have $\sum_{l=1}^{L} T_l$ such vectors as entries in the database

$$\mathbb{S} = \{\{\breve{\mathbf{s}}_{tl}\}_{t=1}^{T_l}\}_{l=1}^{L} . \tag{16}$$

3.2. Objective function

The proposed method inherits the objective function (3), and the contrast function $G(\tilde{\mathbf{y}}_{tn})$ is basically of the form

$$G(\tilde{\mathbf{y}}_{tn}) = \sum_{f=1}^{F} \left(\frac{|y_{ftn}|^2}{v_{ftn}} + \log v_{ftn} \right) . \tag{17}$$

Now, the variance parameters $v_{1tn}, \ldots, v_{Ftn}$ are frequency dependent and constrained in the following manner. Let us define a vector summarizing the variances of all the frequency bins

$$\mathbf{v} = [v_1, \ldots, v_F]^\mathsf{T} . \tag{18}$$

We constrain that the variance vector $\mathbf{v}$ is an entry $\gamma\breve{\mathbf{s}}$ of the database $\mathbb{S}$ with arbitrary scale γ adjustment.

3.3. Variance parameter update by similarity search

To meet the above constraint, we search the database $\mathbb{S}$ for a scale-adjusted entry $\gamma\breve{\mathbf{s}}$ that is most suitable for $\tilde{\mathbf{y}}_{tn}$. Let us define a power-spectrum vector

$$\breve{\mathbf{y}}_{tn} = [|y_{1tn}|^2, \ldots, |y_{Ftn}|^2]^\mathsf{T} \tag{19}$$

of $\tilde{\mathbf{y}}_{tn}$. In this context, the minimization of the contrast function (17) corresponds to the minimization of the Itakura-Saito (IS) divergence $D_{IS}(\breve{\mathbf{y}}_{tn}, \gamma\breve{\mathbf{s}})$ defined by

$$D_{IS}(\breve{\mathbf{y}}_{tn}, \gamma\breve{\mathbf{s}}) = \sum_{f=1}^{F} \left(\frac{|y_{ftn}|^2}{\gamma|s_f|^2} - \log \frac{|y_{ftn}|^2}{\gamma|s_f|^2} - 1 \right) \tag{20}$$

Algorithm 1 Similarity Search-based frequency domain BSS

1: **procedure** $\mathrm{SSBSS}(\{\{\mathbf{x}_{ft}\}_{f=1}^{F}\}_{t=1}^{T})$
2: load database $\mathbb{S}$ prepared as (16)
3: initialize $\mathbf{W}_f$ with an identity matrix for all f
4: **for** $iter = 1$ to $\#iterations$ **do**
5: update $\breve{\mathbf{y}}_{tn}$ for all t, n by (19) and (1)
6: perform similarity search on $\mathbb{S}$ for $\breve{\mathbf{y}}_{tn}$ for all t, n
7: update v_{ftn} for all f, t, n by (21)
8: update $\mathbf{W}_f$ for all f by (8), (10), and (11)
9: **end for**
10: update $\mathbf{y}_{ft}$ for all f, t by (1)
11: **end procedure**

followed by updating the variance parameters

$$\mathbf{v}_{tn} \leftarrow \gamma\breve{\mathbf{s}}_* \ \text{ or } \ v_{ftn} \leftarrow [\gamma\breve{\mathbf{s}}_*]_f \ , f = 1, \ldots, F \tag{21}$$

where $*$ is the index of a database entry that is most similar to $\breve{\mathbf{y}}_{tn}$ in terms of the IS divergence with the scale adjustment

$$\gamma = \frac{1}{F} \sum_{f=1}^{F} \frac{|y_{ftn}|^2}{|s_f|^2} . \tag{22}$$

3.4. Whole algorithm

The proposed method, similarity search-based BSS (SSBSS), is summarized in **Algorithm 1**. Separation matrices $\mathbf{W}_f$ are updated at line 8 by solving the HEAD problem as explained in Sect. 2.1 with v_{tn} in (8) being replaced by v_{ftn}.

4. EXPERIMENTS

We evaluated the performance of the proposed method by experiments. We measured impulse responses from two loudspeakers ($N = 2$) to two microphones ($M = 2$) in a room whose reverberation time was RT_{60} = 200 ms. Then, we made 32 mixture cases by convolving the impulse responses and various combinations of 2-second speech signals. The sampling frequency was 16 kHz. The frame width and shift of STFT were 128 ms and 32 ms, respectively. The separation performance was evaluated in terms of Signal-to-Distortion Ratio (SDR) [25].

Three types of clean source databases consisting of 16 speakers' utterances were prepared. The first one was close database that contains the sources used for making mixtures. Using such a database is in an ideal situation where we verified the basic concept of the proposed method. The second one was open database that did not contain the source time frames used for making mixtures (but contained the same speaker's different utterances). The third one was open+c database where k-means clustering was applied to open database entries per each source ($l = 1, \ldots, L$) and the resultant k-means 10 centroids ($10L$ in total) were added as

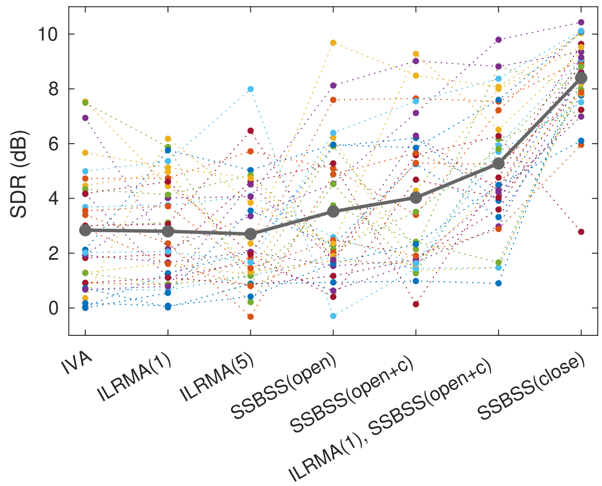

Fig. 2. Separation performance by IVA and ILRMA (existing methods) and four settings of SSBSS (the proposed method). Each dotted line corresponds to a mixture case. The solid grey line represents the average of 32 mixture cases.

new entries. In any cases, the number of database entries was around 30,000 with $F = 1025$ frequency bins.

Figure 2 shows the separation performances. The existing methods, IVA and ILRMA(K) with ranks $K = 1$ and 5, did not attain good SDR values on the average because the mixtures were 2 seconds short. The proposed method SSBSS with close database performed very well. However, the method with open database did not perform well. Adding k-means centroids (open+c) improved the separation a little bit. With these experimental results, we recognized that the database entries should be constructed wisely for better separation. Therefore, we added another type of experiments ILRMA(1), SSBSS(open+c) where ILRMA with rank-1 model $v_{ftn} = b_{fn}a_{tn}$ and the proposed method with open+c database were concatenated. More specifically, ILRMA(1) separated the mixtures at a certain degree and generated new database entries $\check{\mathbf{s}} = [b_{1n}, \ldots, b_{Fn}]^\mathsf{T}$. Then, the proposed method inherited the separation with the dynamically extended database. This combined method improved the separation performances a bit further. However, the gap to the ideal situation SSBSS(close) was still large.

Figure 3 shows typical convergence behavior examples. We observed that $\#iterations = 30$ was sufficient for **Algorithm 1** to converge in terms of not only the objective function (3) but also the IS divergence (20) of each separation to the most similar entry. The execution time was around 20 seconds for a 2-second mixture. We implemented a linear scan similarity search with CUDA C codes executed on a GPU (NVIDIA Tesla V100). All the database entries of the above mentioned size could fit on the GPU memory, and the data load took 1900 ms. The number of queries of the similarity search per iteration (line 7 of **Algorithm 1**) was $N \times T = 2 \times 79 = 158$. One batch similarity search (158 queries for

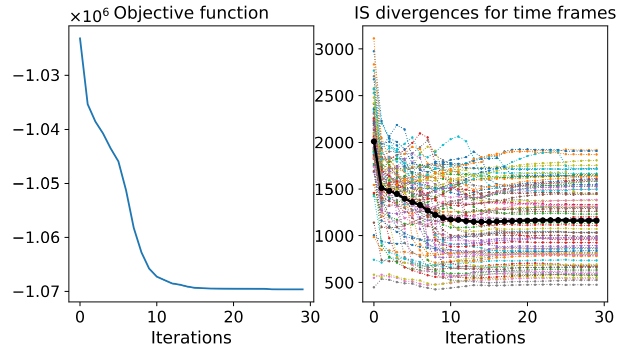

Fig. 3. Convergence behavior examples with close database. Left: objective function (3) with (17). Right: each dotted line represents the IS divergence $D_{IS}(\check{\mathbf{y}}_{t1}, \gamma\check{\mathbf{s}}_*)$ for each time frame $t \in \{1, \ldots, T\}$ of the first $n = 1$ separation. The solid black line represents the average of $T = 76$ frames.

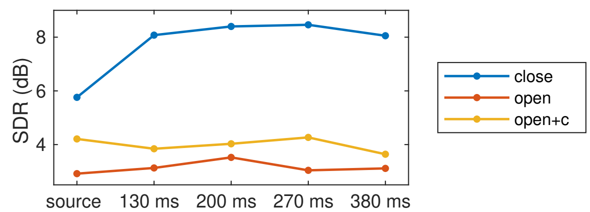

Fig. 4. Averaged separation performance over 32 mixtures of the proposed method by varying the database conditions.

30,000 entries with 1025 dimensionality) took around 230 ms.

Figure 4 shows how reverberant condition mismatches affected the separation performance. The experimental results explained so far were obtained using the databases where dry sources were convolved with the impulse responses whose reverberation time was matched to the mixtures RT_{60} = 200 ms. Here we additionally report the results with databases whose RT_{60} were not 200 ms or where dry sources themselves were used for the entries. We observed that the close cases were largely affected by whether a room reverberant situation was simulated or not (i.e., source). And the difference of reverberation time did not affect much. For the open cases, these conditions were not so important. Rather, a better database should be constructed for better separation.

5. CONCLUSION

The proposed method SSBSS searches a clean database $\mathbb{S}$ for similar entries to current separations $\check{\mathbf{y}}$. The search results are directly used for updating the variance parameters v_{ftn} of LGM-based frequency domain BSS. Experimental results showed that SSBSS with ideal close databases performed very well for 2-second short mixtures. Using open databases lowered the performance considerably. Future work includes developing a method to construct better databases in the open cases. Accelerating the similarity search by approximation [26, 27] is also important to handle a larger database.

6. REFERENCES

[1] S. Haykin, Ed., *Unsupervised Adaptive Filtering (Volume I: Blind Source Separation)*, John Wiley & Sons, 2000.

[2] S. Makino, T.-W. Lee, and H. Sawada, Eds., *Blind Speech Separation*, Springer, 2007.

[3] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, pp. 21–34, 1998.

[4] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001.

[5] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. Speech Audio Processing*, vol. 12, no. 5, pp. 530–538, Sept. 2004.

[6] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. Audio, Speech and Language Processing*, pp. 70–79, Jan. 2007.

[7] A. Hiroe, "Solution of permutation problem in frequency domain ICA using multivariate probability density functions," in *Proc. ICA 2006 (LNCS 3889)*. Mar. 2006, pp. 601–608, Springer.

[8] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. WASPAA*, Oct. 2011, pp. 189–192.

[9] N. Ono, "Auxiliary-function-based independent vector analysis with power of vector-norm type weighting functions," in *Proc. APSIPA ASC*, Dec. 2012, pp. 1–4.

[10] F. Nesta and Z. Koldovskỳ, "Supervised independent vector analysis through pilot dependent components," in *Proc. ICASSP*, 2017, pp. 536–540.

[11] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1626–1641, Sept. 2016.

[12] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, pp. 1–12, 2007.

[13] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. ICASSP*, Mar. 2016, pp. 31–35.

[14] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks.," *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 24, no. 9, pp. 1652–1664, 2016.

[15] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multitalker speech separation," in *Proc. ICASSP*, Mar. 2017, pp. 241–245.

[16] K. Zmolikova, M. Delcroix, K. Kinoshita, T. Higuchi, A. Ogawa, and T. Nakatani, "Speaker-aware neural network based beamformer for speaker extraction in speech mixtures," in *Proc. Interspeech*, 2017.

[17] T. Higuchi, K. Kinoshita, M. Delcroix, K Zmolikova, and T. Nakatani, "Deep clustering-based beamforming for separation with unknown number of sources," in *Proc. Interspeech*, 2017.

[18] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.

[19] S. Mogami, H. Sumino, D. Kitamura, N. Takamune, S. Takamichi, H. Saruwatari, and N. Ono, "Independent deeply learned matrix analysis for multichannel audio source separation," in *Proc. EUSIPCO*, Sept. 2018, pp. 1557–1561.

[20] E. Vincent, S. Arberet, and R. Gribonval, "Underdetermined instantaneous audio source separation via local Gaussian modeling," in *Proc. LVA/ICA*. Springer, 2009, pp. 775–782.

[21] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1118–1133, 2012.

[22] A. Liutkus, J.-L. Durrieu, L. Daudet, and G. Richard, "An overview of informed audio source separation," in *Proc. 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*. IEEE, 2013, pp. 1–4.

[23] S. Amari, S.C. Douglas, A. Cichocki, and H.H. Yang, "Multichannel blind deconvolution and equalization using the natural gradient," in *Proc. IEEE Workshop on Signal Processing Advances in Wireless Communications*, Apr. 1997, pp. 101–104.

[24] A. Yeredor, "On hybrid exact-approximate joint diagonalization," in *Proc. IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 2009, pp. 312–315.

[25] E. Vincent, S. Araki, F. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, V. Gowreesunker, D. Lutter, and N.Q.K. Duong, "The signal separation evaluation campaign (2007–2010): Achievements and remaining challenges," *Signal Processing*, vol. 92, no. 8, pp. 1928–1936, 2012.

[26] K. Aoyama, S. Watanabe, H. Sawada, Y. Minami, N. Ueda, and K. Saito, "Fast similarity search on a large speech data set with neighborhood graph indexing," in *Proc. ICASSP*, 2010, pp. 5358–5361.

[27] K. Aoyama, K. Saito, H. Sawada, and N. Ueda, "Fast approximate similarity search based on degree-reduced neighborhood graphs," in *Proc. 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, pp. 1055–1063.