Journal of the Korea Academia-Industrial cooperation Society
Vol. 22, No. 10 pp. 544-551, 2021

https://doi.org/10.5762/KAIS.2021.22.10.544
ISSN 1975-4701 / eISSN 2288-4688

Parameter Estimation of Source Image Using Least-Squares for Dual Channel Underdetermined Convolutive Blind Speech Separation

Jounghoon Beh
Intelligent Signal Processing Lab., Korea University

이중 채널에서 블라인드 음성 분리를 위한 최소 제곱법 기반의 파라미터 추정 방법

배정훈
고려대학교 지능신호처리연구실

Abstract Blind source separation for the cocktail-party problem has long been a challenge in Artificial Intelligence (AI) for military applications. The cocktail-party problem states that an intelligent machine cannot locate and listen to the target speaker's voice when it is surrounded by other sound sources such as in a cocktail party. This paper proposes a novel method to estimate mixing parameters of underdetermined and convolutive mixture of speech sources given a dual-channel microphone array. In the proposed method, the optimal estimate of mixing parameters and the source image is obtained using the least-squares principle. As a result, the scaling ambiguity, which is common in conventional blind source separation methods, is alleviated. Performance evaluation of the proposed method is conducted with a public dataset, SiSEC 2008, and experimental results show the present method's validity in terms of various SNR measures compared to other state-of-the-art techniques in the field.

요 약 칵테일 파티 문제를 해결하기 위한 일환으로서 블라인드 음원 분리 기술은 군사용 인공지능 활용 분야에서 오랫동안 연구되어왔다. 칵테일 파티 문제는 언어이해를 목적으로 하는 지능 시스템이 다양한 음원이 존재하는 상황에서 타겟사용자의 목소리를 분간해 내지 못하여 사용자 명령을 이해하지 못하는 현상을 말한다. 본 논문에서는 이중 채널 마이크로폰 어레이 상황에서 언더디터민드 (underdetermined) 하고 합성곱 형식으로 섞여진 신호의 혼합계수 (mixing parameter)를 추정하는 새로운 방법을 제안한다. 최소제곱법 원리를 이용하여 혼합계수들을 추정하는 방법이 소개되며, 분리된 신호는 마이크로폰에서의 신호 형상 (signal image) 레벨로 구해진다. 이 결과, 일반적으로 블라인드 음원 분리 기술에서 맞닥뜨리는 스케일링 모호성 (scaling ambiguity)이 없어지게 된다. 제안한 방법은 최소제곱법을 기반을 두며, 일반적으로 블라인드 음원 분리 기술에서 맞닥뜨리는 스케일링 모호성 (scaling ambiguity)를 없앤다. 제안한 방법의 성능은 학계에 공개된 SiSEC 2008 데이터셀을 이용해 측정되었으며, 다양한 종류의 신호 대 잡음비를 측정하고, 그 측정값을 같은 분야의 최신 기술들과 비교함으로써 그 타당성을 입증하였다.

Keywords : Artificial Intelligence, Blind Source Separation, Least Square Estimation, Military Application, Speech Enhancement

*Corresponding Author : Jounghoon Beh(Korea Univ.)
email: jounghoon.beh@gmail.com

Received September 13, 2021 Revised September 27, 2021
Accepted October 1, 2021 Published October 31, 2021

1. Introduction

Speech technologies for enabling natural language interfaces have been emerging in the military AI applications[1]. Among them, the blind speech separation (BSS) has been one of the challenging areas in the field of military signal processing[2]. A number of BSS techniques have shown their effectiveness when there is speech sparseness in the Time-Frequency (TF) domain[3,4]. In a reverberant environment, BSS turns into a so-called "convolutive mixture" problem, and the separation is usually performed narrowband-wise in the frequency domain. The approaches for the frequency domain BSS include i) spectral bin clustering[4,5] or mixing parameter estimation[6]; ii) soft/binary masking [3,4,7-12] or separation based on l_p norm minimization[13]; iii) permutation alignment[4,14]; iv) the Deep Neural Network (DNN) based monaural speech separation[11,15,16]; v) the DNN based multi-channel speech separation[17-19].

As the DNN has shown improved performance in various machine learning tasks, the DNN based source separation technologies also have emerged. Heymann *et al.*[10] proposed a method of generating a binary mask in the magnitude Short-Time Fourier Transform (STFT) domain using stacks of fully-connected layers to the Bidirectional Long Short-Term Memory (BLSTM). Williamson *et al.*[11] expanded the domain of mask generation to the complex STFT domain. Erdogan *et al.*[12] incorporated a phase difference between speech and noisy speech in the mask generation process in the magnitude STFT domain.

It should be noted that the DNN based separation method suffers generalization problems since the environment where its training data are collected may not match with the testing environment. Such mismatches occur from difference in microphone array geometry, type of noise, range of signal-to-noise ratio, and human factors (age, gender, accent) in training audio data. In the traditional approach of speech separation, however, such a generalization problem is relatively eased because it usually models statistical characteristics of speech with parametric probability models, and tries to figure out how to accurately estimate its parameter. Those modeling assumptions on speech signals are not confined to a human factor or type of noise. Therefore, in such an aspect, a traditional approach is more effective to mitigate generalization problems over recent DNN approaches.

In this paper, we developed a simple but effective spectral bin clustering method for identifying the speech source. Note that the term "spectral bin" denotes the time-frequency bin of a spectrogram which is derived from the STFT. In the speech sparseness based approaches, the spectral bin clustering is mostly affective to overall performance of the separation system. It is because the mixing parameter are estimated from the clusters, its estimate is fed into the separation process such as l_p norm minimization and permutation alignment. Moreover, correctly clustered spectral bin itself can serve as a separated speech source[3] or at least the mixing parameters are easily estimated from the clustered spectral bin[6].

In the literature, various methods for spectral bin clustering have been proposed using weighted histogram[3], mixture of Gaussian[4], and mixture of wrapped-Gaussian[8]. However, strictly speaking, all those approaches are optimal in clustering signal ratio between two microphones rather than clustering the source components themselves. In that case, the estimation might be inaccurate due to some abnormal outlier value of its ratio which is inherently resulted from its division form. Winter *et al.*[6] tackled this problem earlier by employing a hierarchical clustering algorithm to the spectral bin clustering to ease the outlier problem. However, their method inherently required determining the a Priori parameters, which bring some ambiguity in their implementation. We propose a novel method to directly estimate the mixing parameter with data under the signal sparseness assumption[3]. Other than that, no additional assumptions on a probability distribution on the signal modeling are required.

The remainder of this paper is structured as follows. In Section 2 we build a signal model for deriving the proposed method. In Section 3 we present the main idea of this work. Performance evaluation is shown in Section 4. Section 5 concludes this paper.

2. Signal modeling

Suppose that we have N sources of time-domain signal $s_n(t), n=1,\ldots,N$ distributed around a microphone pair. Those source signals are convolutively mixed and captured by each microphone as

$$x_m(t) = \sum_{n=1}^{N}\sum_{l} h_{mn}(l)s_n(t-l), m=1,2 \qquad (1)$$

where $h_{nm}(l)$ is the impulse response from n^{th} source to m^{th} microphone. Then time-domain signal of m^{th} microphone, $x_m(t)$, is transformed to time-frequency domain via the STFT, and leads to

$$X_m(\tau,f) = \sum_{n=1}^{N} H_{mn}(f)S_n(\tau,f),\ m=1,2 \qquad (2)$$

where τ and f denotes the frame and frequency index, respectively.

They are formed in a vector as follows:

$$\mathrm{x}(\tau,f) = \begin{bmatrix} X_1(\tau,f) \\ X_2(\tau,f) \end{bmatrix} \qquad (3)$$
$$= \begin{bmatrix} H_{11}(f) & \cdots & H_{1N}(f) \\ H_{21}(f) & \cdots & H_{2N}(f) \end{bmatrix} \begin{bmatrix} S_1(\tau,f) \\ \vdots \\ S_N(\tau,f) \end{bmatrix}.$$

where $H_{mn}(f)$ and $S_n(\tau,f)$ represent mixing parameter and source signal of frequency f at τ^{th} frame , respectively.

Source speech image at each microphone is formed as

$$\mathrm{i}_m(\tau,f) = [H_{m1}(f)S_1(\tau,f),\cdots,H_{mN}(f)S_N(\tau,f)]^T, \qquad (4)$$
$$m=1,2$$

Using (3) the input signal $\mathrm{x}(\tau,f)$ can be represented with the source speech image as follows:

$$\mathrm{x}(\tau,f) = \mathrm{R}_m(f)\mathrm{i}_m(\tau,f), \quad m=1,2 \qquad (5)$$

where $\mathrm{R}_m(f)$ is the mixing parameter matrix for the source image.

Let $R_{12n}(f) = H_{1n}(f)/H_{2n}(f)$ and $R_{21n}(f) = H_{2n}(f)/H_{1n}(f)$, n =1,$\cdots$,N and then (5) is rewritten as

$$\mathrm{x}(\tau,f) = \mathrm{R}_1(f)\mathrm{i}_1(\tau,f) \qquad (6)$$
$$= \begin{bmatrix} 1 & \cdots & 1 \\ R_{211}(f) & \cdots & R_{21N}(f) \end{bmatrix} \begin{bmatrix} H_{11}(f)S_1(\tau,f) \\ \vdots \\ H_{11}(f)S_1(\tau,f) \end{bmatrix},$$

and equivalently

$$\mathrm{x}(\tau,f) = \mathrm{R}_2(f)\mathrm{i}_2(\tau,f) \qquad (7)$$
$$= \begin{bmatrix} R_{121}(f) & \cdots & R_{12N}(f) \\ 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} H_{21}(f)S_1(\tau,f) \\ \vdots \\ H_{21}(f)S_1(\tau,f) \end{bmatrix},$$

Generally conventional methods attempt to solve (3) directly in terms of the mixing parameter $H_{mn}(f)$ and recover the signal source $S_n(\tau,f)$. However, this approach inevitably encounters a scaling ambiguity problem, and additional post-processing is required to resolve the problem. Our approach is to recover the signal image, $H_{mn}(f) \cdot S_n(t,f)$, not the signal source $S_n(\tau,f)$ itself. By this way, the scaling ambiguity can be inherently avoided.

The performance is evaluated on the task of speech enhancement assuming that the properly estimated mixing parameter leads to a good quality of recovered speech signal.

3. Proposed method

3.1 Mixing parameter estimation

We estimate the transfer function ratio,

$R_{12n}(f)$ and $R_{21n}(f)$, by minimizing

$$E[|X_1(\tau,f)-R_{12n}(f)X_2(\tau,f)|^2 + E[|X_2(\tau,f)-R_{21n}(f)X_1(\tau,f)|^2,\ n=1,2 \tag{8}$$

In order to calculate (8), we first need to determine which source the input signal $\mathrm{x}(\tau,f)$ belongs to, and then the signals are clustered in terms of its originated sources. Using the clustered data, $R_{12n}(f)$ and $R_{21n}(f)$ are updated. These two procedures are repeated until the value of $R_{12n}(f)$ and $R_{21n}(f)$ are converged.

Let's assume that T observations at frequency f, namely $\mathrm{x}(1,f),\ldots,\mathrm{x}(T,f)$ are given. The goal of the proposed method is to classify those observations into corresponding N source sets. We define $\xi_n(f)$ as a set of $\mathrm{x}(\tau,f)$ hat mostly comprises n^{th} source $S(\tau,f)$:

$$\xi_n(f)=\{\mathrm{x}(\tau,f)\,|\,C(\mathrm{x}(\tau,f),\mathrm{r}_n(f))<C(\mathrm{x}(\tau,f),\mathrm{r}_k(f)),\ n\neq k, 1\leq n\leq N, 1\leq k\leq N\} \tag{9}$$

where $\mathrm{x}(\tau,f)=[X_1(\tau,f)\ \ X_2(\tau,f)]^T$ and $\mathrm{r}(f)=[R_{12n}(\tau,f)\ \ R_{21n}(\tau,f)]^T$ and $\mathrm{r}_n(f)$ is obtained by minimizing the summation of cost function for each source n:

$$\mathrm{r}_n(f)=\underset{\mathrm{r}_n(f)}{\operatorname{argmin}}\sum_{\mathrm{x}(\tau,f)\in\xi_n(f)} C(\mathrm{x}(\tau,f),\mathrm{r_n}(\mathrm{f})). \tag{10}$$

A cost function $C(\mathrm{x}(\tau,f),\mathrm{r}_n(f))$ is defined as a similarity between $X_1(\tau,f)$ and $X_2(\tau,f)$ given $\mathrm{r}_n(f)$:

$$C(\mathrm{x}(\tau,f),\mathrm{r}_n(f))=|X_1(\tau,f)-R_{12n}(f)X_2(\tau,f)|^2 + |X_2(\tau,f)-R_{21n}(f)X_1(\tau,f)|^2 ,n=1,\ldots,N \tag{11}$$

Now the overall criterion for clustering given observations $\mathrm{x}(\tau,f),\tau=1,\ldots T$ into each source set $\xi_n(f),n=1,\ldots,N$ can be written as

$$\xi^*(f)=\underset{\xi(f)}{\operatorname{argmin}}\sum_{n=1}^{N}\sum_{\mathrm{x}(\tau,f)\in\xi_n(f)} C(\mathrm{x}(\tau,f),\mathrm{r}_n(f)),\quad \tau=1,\ldots,T \tag{12}$$

In practice, the $\xi^*(f)$ is obtained in an iterative manner as follows. At first, elements of a set $\xi_n(f),n=1,\ldots N$ are determined for all N sources using (9) with the previous estimate of $\mathrm{r}_n(f)$. Then, using (10), a new estimate of $\mathrm{r}_n(f)$ is obtained with those $\xi_n(f)$ which is determined by (9).

3.2 Source separation and recovery

Since the proposed algorithm is designed to estimated appropriate mixing parameter, we adopted ground-truth permutation information. After the permutation, a binary mask was generated and applied to separate source images at each microphone as follow[3]:

$$M_n(\tau,f)=\begin{cases}1 & \text{if } \mathrm{x}(\tau,f)\in\xi_n(f),\\ 0 & otherwise\end{cases} \tag{13}$$

The mask value is multiplied to input signal in the time-frequency domain. The input signal is then recovered by inverse STFT to time-domain signal. The transformed data are merged in overlap-and-add manner.

4. Experiments

4.1 Task and settings

The dataset and measurement are provided from SiSEC speech separation campaign[20] which is available for the public use of BSS algorithm evaluation. Data of SiSEC 2008 database are recorded with 16 kHz sampling rate. Among the data, we picked up the audio data that was lively recorded in a reverberant environment where the reverberation time was either T_{60}=130 ms or 250 ms. The data recorded in the T_{60}=130 ms environments are transformed to STFT domain with the Hanning window of 2,048 samples and 512 samples overlapping period. For the data recorded in the T_{60}=250 ms, we set the window size to be 4,096 samples. The spacing between microphones was set to 5cm or 1m. The number of speakers are 3 or 4. Therefore there are 8 combinations of data settings. (microphone spacing / reverberation time / number of source)

4.2 Performance measure

Four objective evaluation scores are measured per data setting. Those are the Signal to Distortion Ratio (SDR), the image Signal to Spatial Distortion Ratio (ISR), the Signal to Interference Ratio (SIR), and Signal to Artifacts Ratio (SAR). The detailed derivation of these scores can be found in [21]. They are scaled as dB, and larger numbers mean better performance. The SDR indicates power ratio between original signal and overall distortion components due to imperfection of the BSS algorithm. The ISR represents how much stronger the energy of target speech over the amount of damaged part of the original signal after the separation. The SIR accounts for how effectively interference (other speech sources) are removed. The SAR measures the energy ratio between the original signal over the unwanted artifact signal generated by the algorithm.

4.3 Results

We compared performance of the proposed method with four other state-of-the-art BSS technologies that also have evaluated their performance with the same SiSEC 2008 data set. The first method, denoted as “Ozerov”[22], estimates mixing parameters using the expectation-maximization (EM) algorithm[23]. The second one, denoted as “Nesta^1”[24], estimates mixing parameters with the natural gradient algorithm. The third method, denoted as “Nesta^2” is all the same with [24] but they added an additional Wiener post-processing to recover the original speech signal. The last method, denoted as “Iso”[25], is a combination of benefits from two different algorithms: [4] and [26]. A full-rank method[26] was adopted to estimate the source image, and for the initialization of the center value of the spectral bin cluster, the bin-wise clustering method[4] was adopted. The “IBM” in the table denotes the results from signal separation using the ideal binary mask provided by the dataset publisher for benchmarking and it shows upper-bounds on the performance of the binary masking-based method[27]. We have taken and compared the performance result of other methods from [28] where it has been open to public.

4.3.1 Blind source separation with three speech sources

In Table 1, we summarized and compared the performance of blind source separation for the test case where three speech signal sources exist, the distance between microphones is 5 cm long, and the reverberation time is 130 ms. The proposed method outperforms the other state-of-the-art BSS algorithm in terms of the ISR and the SIR. It means the proposed method successfully clustered spectral bins according to its source. However, it generated unwanted musical noise so the SAR score of the proposed method ranked lowest, so even though the interferences are well separated out, the SDR could not show higher value than the others.

Table 1. Performance comparisons of three speech source separation where reverberation time is 130ms

Mic. spacing	Measure	Method					
		Proposed	Ozerov	Nesta^1	Nesta^2	Iso	IBM
5cm	SDR	6.7	4.6	6.4	6.7	**7.1**	6.4
	ISR	**12.5**	9.2	10.2	12.4	11.4	10.2
	SIR	**13.1**	8.4	10.4	12.0	10.5	10.4
	SAR	8.0	8.4	10.1	9.4	**10.8**	10.1
1m	SDR	7.3	3.0	6.2	**7.4**	4.6	10.0
	ISR	**13.2**	7.6	8.8	12.1	8.3	18.6
	SIR	**13.7**	4.8	10.5	12.3	7.0	20.2
	SAR	8.4	7.3	9.8	**9.9**	9.0	10.6

For the next evaluation, we changed the testing acoustic environments harsher by setting the reverberation time longer (130ms → 250ms) and the evaluation results are presented in Table 2. It makes the phase difference spread widely in the phase axis all over the frequency band. It affects badly for its performance to cluster and separate each spectral bin into its own speech sources. Therefore, the overall performance is expected to be lower than the one measured when it is 130ms (Table 1). For the 1 m case, the proposed method achieved the best performance in the SDR thanks to higher score obtained in the ISR and SIR field. It should be noted that the proposed methods showed a robustness against the reverberation.

Table 2. Performance comparisons of three speech source separation where reverberation time is 250ms

Mic. spacing	Measure	Method					
		Proposed	Ozerov	Nesta^1	Nesta^2	Iso	IBM
5cm	SDR	6.6	4.3	6.0	**6.9**	6.1	10.1
	ISR	**12.3**	9.0	9.8	11.2	10.4	18.8
	SIR	**12.6**	8.5	10.0	11.6	9.2	20.2
	SAR	7.5	8.4	8.9	9.6	**10.2**	10.7
1m	SDR	**6.9**	4.8	4.6	6.1	5.8	9.6
	ISR	**12.6**	9.3	7.0	10.2	9.6	18.0
	SIR	**12.9**	8.6	8.5	10.4	8.6	19.5
	SAR	8.1	8.1	8.0	9.4	**10.0**	10.1

4.3.2 Blind source separation with four speech sources.

We evaluated and compared the performance of the BSS methods in harsher condition by adding one more speech source in the mixture signal. The other conditions remained the same as we conducted in Section 4.3.1. Note that we compared the proposed method with the other technologies but excluded the "Iso" method because it was not reported in public.

In Table 3, we presented performance scores for each algorithm in the case where reverberation time is 130 ms. Compared to Table 1, regardless of method, every performance was degraded due to increment of number of speech sources to be separated. Since the number of speech sources increased, it is highly likely that in a time-frequency bin, multiple speech sources are overlapped. Therefore, it is harder to meet the signal sparseness assumption[3], which is primarily required, for all methods to work, than the three speech sources case.

Table 3. Performance comparisons of four speech source separation where reverberation time is 130 ms

Mic. spacing	Measure	Method				
		Proposed	Ozerov	Nesta^1	Nesta^2	IBM
5cm	SDR	**4.8**	2.7	4.3	4.4	4.3
	ISR	**9.6**	6.9	7.4	8.3	7.5
	SIR	**10.2**	5.7	7.9	9.0	8.0
	SAR	5.9	6.5	**7.1**	6.8	7.1
1m	SDR	**5.1**	2.7	3.5	3.9	7.7
	ISR	**10.1**	6.5	6.0	7.7	15.3
	SIR	**10.6**	4.7	5.7	6.9	17.2
	SAR	5.9	5.9	**6.5**	6.3	8.1

Table 4 shows the result of performance where the same condition is remained as the same as Table 3 but the reverberation time is increased to 250 ms. Throughout all the experimental settings so far, this is the most challenging acoustic condition for the algorithms to achieve proper separation of the mixed speech sources. As we compared to Table III, it has shown that the performance of all algorithm in all aspects are even more worsened. The proposed method outperformed the other method in all scores but the SAR

Table 4. Performance comparisons of four speech source separation where reverberation time is 250 ms

Mic. spacing	Measure	Method				
		Proposed	Ozerov	Nesta^1	Nesta^2	IBM
5cm	SDR	**4.2**	2.4	2.6	3.1	8.5
	ISR	**8.2**	6.5	5.9	6.7	17.0
	SIR	**8.4**	4.5	4.4	5.5	18.9
	SAR	5.0	5.5	5.4	**6.0**	8.7
1m	SDR	**5.0**	1.8	3.2	4.1	8.6
	ISR	**9.6**	5.5	5.5	7.6	16.4
	SIR	**10.0**	3.1	5.9	7.2	18.7
	SAR	5.7	5.0	5.6	**6.7**	8.9

Throughout the overall results given in Table 1 to Table 4, the proposed method has shown its robust performance in various condition but especially in the condition of four speech sources.

5. Conclusion

A novel method of estimating mixing parameters is proposed for blind speech source separation in the case of a two-channel microphone array. The signal sparseness assumption is employed to model the speech source mixture. Unlike the conventional approaches, we directly estimate optimal mixing parameters in terms of least squares principle rather than signal ratio. In the experimental results, with the ideal permutation alignment, the proposed method has shown improved performance compared to those of the state-of-the-art techniques. Future work may include extending the proposed scheme from two-channel to arbitrary number of channels, and devising an efficient permutation alignment algorithm in order to incorporate it with the proposed mixing parameter estimation algorithm.

References

[1] F. E. Morgan, B. Boudreaux, A. J. Lohn, C. Curriden, K. Klima and D. Grossman, "Military applications of artificial intelligence: ethical concerns in an uncertain world," RAND PROJECT AIR FORCE SANTA MONICA CA SANTA MONICA, 2020.

[2] N. Doukas and N. V. Karadimas, "A blind source separation based cryptography scheme for mobile military communication applications," WSEAS Trans. Commun, vol. 7, no. 12, pp. 1235-1245, 2008.

[3] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," IEEE Trans. Signal Process., vol. 52, no. 7, pp. 1830-1847, 2004.
DOI: http://dx.doi.org/10.1109/TSP.2004.828896

[4] H. Sawada, S. Araki and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," IEEE Audio, Speech, Language Process., vol. 19, no. 3, pp. 516-527, 2010.
DOI: http://dx.doi.org/10.1109/TASL.2010.2051355

[5] B. Loesch and B. Yang, "Blind source separation based on time-frequency sparseness in the presence of spatial aliasing," in LVA/ICA 2010, 2010.
DOI: http://dx.doi.org/10.1007/978-3-642-15995-4_1

[6] S. Winter, W. Kellermann, H. Sawada and S. Makino, "MAP-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and l1-norm minimization," EURASIP Journal on Advances in Signal Processing, vol. 2007, pp. 1-12, 2007.
DOI: http://dx.doi.org/10.1155/2007/24717

[7] Y. Izumi, N. Ono and S. Sagayama, "Sparseness based 2ch BSS using the EM algorithm in reverberant environment," in WASPAA 2007, 2007.
DOI: http://dx.doi.org/10.1109/ASPAA.2007.4393015

[8] S. Araki, T. Nakatani, H. Sawada and S. Makino, "Stereo source separation and source counting with MAP estimation with Dirichlet prior considering spatial aliasing problem," in ICA 2009, 2009.
DOI: http://dx.doi.org/10.1007/978-3-642-00599-2_93

[9] M. Cobos and J. J. Lopez, "Maximum a posteriori binary mask Estimation for underdetermined source separation using smoothed posteriors," IEEE Audio, Speech, Language Process., vol. 20, no. 7, pp. 2059-2064, 2012.
DOI: http://dx.doi.org/10.1109/TASL.2012.2195654

[10] J. Heymann, L. Drude and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in ICASSP 2016, 2016.
DOI: http://dx.doi.org/10.1109/ICASSP.2016.7471664

[11] D. S. Williamson, Y. Wang and D. Wang, "Complex Ratio Masking for Monaural Speech," IEEE/ACM Transactions On Audio, Speech, And Language Processing, vol. 24, no. 3, pp. 483-492, 2016.
DOI: https://doi.org/10.1109/TASLP.2015.2512042

[12] H. Erdogan, J. Hershey, S. Watanabe and J. L. Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in ICASSP 2015, 2015.
DOI: https://doi.org/10.1109/ICASSP.2015.7178061

[13] E. Vincent, "Complex nonconvex lp norm minimization for underdetermined source separation," in ICA 2001, 2001.
DOI: https://doi.org/10.1007/978-3-540-74494-8_54

[14] H. Sawada, S. Araki, R. Mukai and S. Makino, "Grouping separated frequency components by estimating propagation model parameter in frequency-domain blind source separation," IEEE Audio, Speech, Language Process., vol. 15, no. 5, pp. 1592-1604, 2007.
DOI: https://doi.org/10.1109/TASL.2007.899218

[15] E. M. Grais, M. U. Sen and H. Erdogan, "Deep neural networks for single channel source separation," in ICASSP, 2014.
DOI: https://doi.org/10.1109/ICASSP.2014.6854299

[16] P.-S. Huang, M. Kim, M. Hasegawa-Johnson and P. Smaragdis, "Deep learning for monaural speech separation," in ICASSP 2014, 2014.
DOI: https://doi.org/10.1109/ICASSP.2014.6853860

[17] Y. Jiang, D. L. Wang, R. S. Liu and Z. M. Feng, "Binaural classification for reverberant speech segregation using deep neural networks," IEEE/ACM Trans. Audio Speech Lang. Proc., vol. 22, pp.

2112-2121, 2014.
DOI: https://doi.org/10.1109/TASLP.2014.2361023

[18] Y. Yu, W. Wang and P. Han, "Localization based stereo speech source separation using probabilistic time-frequency masking and deep neural networks," EURASIP J. Audio Speech Music Proc., vol. 2016, pp. 1-18, 2016.
DOI: http://dx.doi.org/10.1186/s13636-016-0085-x

[19] X. Zhang and D. L. Wang, "Deep learning based binaural speech separation in reverberant environments," IEEE/ACM Trans. Audio Speech Lang. Proc., vol. 25, pp. 1075-1084, 2017.
DOI: https://doi.org/10.1109/TASLP.2017.2687104

[20] [Online]. Available: http://sisec.wiki.irisa.fr

[21] V. Emiya, E. Vincent, N. Harlander and V. Hohmann, "Subjective and objective quality assessment of audio source separation," IEEE Trans. On Audio, Speech and Language Process., vol. 19, no. 7, pp. 2046-2057, 2011.
DOI: https://doi.org/10.1109/TASL.2011.2109381

[22] A. Ozerov, E. Vincent and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," IEEE Trans. on Audio, Speech and Language Process., vol. 20, no. 4, pp. 1118-1133, 2011.
DOI: https://doi.org/10.1109/TASL.2011.2172425

[23] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society: Series B (Methodological), vol. 39, no. 1, pp. 1-22, 1977.
DOI: https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

[24] F. Nesta and M. Omologo, "Convolutive underdetermined source separation through weighted interleaved ICA and spatio-temporal source correlation," in LVA/ICA 2012, 2012.
DOI: https://doi.org/10.1007/978-3-642-28551-6_28

[25] K. Iso, S. Araki, S. Makino, T. Nakatani, Y. Yamada, T. Yamada and A. Nakamura, "Blind source separation of mixed speech in a high reverberation environment," in HSCMA2011, 2011.
DOI: https://doi.org/10.1109/HSCMA.2011.5942406

[26] N. Q. K. Duong, E. Vincent and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," IEEE Trans. on Audio, Speech and Language Process., vol. 18, no. 7, pp. 1830-1840, 2010.
DOI: https://doi.org/10.1109/TASL.2010.2050716

[27] S. Araki, *et al.*, "The 2011 signal separation evaluation campaign (SiSEC2011):-audio source separation." In International Conference on Latent Variable Analysis and Signal Separation, pp. 414-422. Springer, Berlin, Heidelberg, 2012.
DOI: https://doi.org/10.1007/978-3-642-28551-6_51

[28] [Online]. Available: https://www.irisa.fr/metiss/SiSEC11/underdetermined/underdetermined_test_all.html

Jounghoon Beh [Regular member]

- Feb. 2001 : Korea Univ., EERE, BS
- Feb. 2003 : Korea Univ., EE, MS
- Aug. 2008 : Korea Univ., ECE, PhD
- Sep. 2021 ~ current : Korea Univ., ISPL, Research Professor

〈Research Interests〉

AI, Machine Learning, Speech Signal Processing