Journal of the Korea Academia-Industrial cooperation Society
Vol. 22, No. 10 pp. 752-759, 2021

https://doi.org/10.5762/KAIS.2021.22.10.752
ISSN 1975-4701 / eISSN 2288-4688

MAP Adaptation with Mixtures of von Mises Distributions and Its Application to Underdetermined Convolutive Blind Source Separation

Jounghoon Beh
Intelligent Signal Processing Lab., Korea University

혼합 폰 미제스 확률 분포에 대한 최대 사후 확률 기반 적응 기법과 블라인드 음원 분리에의 적용

배정훈
고려대학교 지능신호처리연구실

Abstract In this paper, we propose a Maximum a Posteriori (MAP) adaptation method for the parameter of Phase Difference of Arrival (PDOA) distribution to classify sound sources in the time-frequency domain. The PDOA is clustered via Mixtures of von Mises distribution (MovM) which is efficacious in modeling a 2 circular domain. The proposed method is employed to the task of clustering a spectral bin in terms of a sound source, which is a crucial part of the mask-based blind source separation. To robustly estimate the model parameter of MovM of the PDOA, we first build an incidence angle distribution and then adapt the parameters to each frequency band. The clustered spectral bin is used to build a time-frequency (TF) mask to separate the mixed audio signal based on the signal sparseness property.

요 약 본 논문에서는 시간-주파수 영역에서 음원분리를 위해 도달음원의 위상차 분포 추정을 위한 최대 사후 확률 기반의 파라미터 적응기법을 제안한다. 도달 음원의 위상차 데이터는 혼합 폰 미제스 분포로 가정하여 음원의 각도 위치에 따라 분류되었으며, 이 혼합 폰 미제스 분포는 위상차와 같은 2π 순환 데이터를 모델링할 때 효과적이다. 제안한 방법은 혼합 음원의 스펙트로그램 상에서 시간 및 주파수 구간을 각각의 음원으로 클러스터링하는 용도에 사용되었으며, 이는 마스크 기반 블라인드 음원 분리기술에서 가장 핵심적인 부분이다. 혼합 폰 미제스 분포로 모델링 된 도달음원 위상차 분포의 파라미터들을 안정적으로 추정하기 위하여, 먼저 각 도착 신호의 각도 분포를 구성하였으며, 이 각도 분포를 바탕으로 각 주파수 밴드의 위상차 분포를 나타내는 파라미터들을 각각의 주파수 대역 별로 추정하였다. 이를 바탕으로 스펙트로그램상에서 음원별로 시간-주파수 영역을 구분하고 이 영역들은 시간-주파수 마스크를 생성하는데 쓰여졌다. 제안한 방법은 SiSEC 2008 데이터셑을 이용하여 다양한 신호 대 잡음비 값으로 평가되었으며, 그 타당성이 입증되었다.

Keywords : Artificial Intelligence, Military Application, Von Mises Distribution, Maximum A Posteriori, Blind Source Separation

*Corresponding Author : Jounghoon Beh(Korea Univ.)
email: jounghoon.beh@gmail.com
Received September 15, 2021 Revised September 28, 2021
Accepted October 1, 2021 Published October 31, 2021

1. Introduction

The task of blind source separation (BSS) is to estimate source signals in the presence of multiple sources. Application of the BSS for speech includes pre-processing of automatic speech recognition for military applications[1] and mobile military communications[2]. Recently the signal sparseness-based separation techniques have emerged and attracted much attention. These techniques can be categorized into 1) Time-Frequency (TF) masking[3-6] and 2) mixing parameter estimation[7,8]. However, the common prerequisite for those techniques is clustering the spectral bin of the input signal into each sound source in an unsupervised manner. In many kinds of literatures, the Time Difference of Arrival (TDOA)[3], signal orientation[6], or Phase Difference of Arrival (PDOA)[5,9], are incorporated with a Bayesian framework, and Expectation-Maximization (EM) is popularly used to learn parameters of the cluster distribution. However, the measured TDOA above the spatial aliasing frequency suffers 2π-ambiguity problem, and therefore they are not reliable to be used. The 2π-ambiguity problem means that given a TDOA value measured from the phase difference, we cannot explicitly determine the direction of sound source because the phase difference is observed in a 2π- modulo manner, i.e. $2\pi k, k = 0, \pm 1, \pm 2, \cdots$.

A number of approaches to avoid the 2π -ambiguity problem of TDOA have been introduced recently. In some works, the 2π -ambiguity has been handled in a probabilistic manner by setting it as a model parameter of a PDOA distribution[4,9]. Izumi *et al.*[4] has used TDOA as a model parameter of the complex Gaussian Mixture Model (GMM) which represents the distribution of the input speech signal in a complex time-frequency domain.

Recently, the Deep Neural Network (DNN) has been showing improved performance on various types of speech separation tasks: time-frequency masking estimation for speech enhancement [10-12] monaural source separation[11-14], and multi-channel source separation[15-17]. For such a learning-based speech separation, it is hard to avoid inherent performance degradation when generalization fails. It is mainly due to a mismatch between training data and testing ones. Many conditions should be considered when collecting (or selecting) training sets in terms of environments where the algorithm is actually applied: array geometry, type of noise, range of signal-to-noise ratio, speakers. On the other hand, the traditional approach of speech separation adopts a proper probabilistic model of the speech signal and estimates its model parameter based on statistical criteria[18,19].

The goal of this paper is to develop an efficacious probabilistic model for representing the PDOA distribution in order to cluster PDOA observation in the spectral bin, especially in the case of a two-channel microphone array. The MovM is adopted to build the distribution, and the spectral bins are clustered in terms of a probabilistic score of PDOA. The remainder of this paper is organized as follows. In Section 2 our signal modeling is introduced. Based on that, in Section 3, we propose a MAP adaptation method to estimate the parameter of PDOA distribution to classify sound sources without concerning 2π-ambiguigity problem. In Section 4 performance of the proposed method is evaluated in terms of speech enhancement. Concluding remarks are presented in Section 5.

2. Signal modeling

Suppose that we have N_S sources of signal s_i, $i = 1, \ldots, N_S$ distributed around two microphones. The signals are convolutively mixed and captured by each microphone as

$$x_j(t)=\sum_{i=1}^{N_s}\sum_{l}h_{ji}(l)s_i(t-l)+n(t),\ j=1,2 \quad (1)$$

where j is the microphone index, $h_{ji}(l)$ is the impulse response from the i^{th} source to the j^{th} microphone, and $n(t)$ is ambient noise. The signals are converted to TF domain via the Short-Time Fourier Transform (STFT), and leads to

$$X_j(\tau,\omega)=\sum_{n=1}^{N}H_{ji}(f)S_i(\tau,\omega)+N(\tau,\omega),\ m=1,2 \quad (2)$$

where τ denotes a time index ω denotes an angular frequency.

With an assumption that the signal sources are not moving, the phase difference between signals of two microphones can represent where the signal source they are located, and the PDOA at is formed as:

$$\theta_{\pi\omega}=\arg\left[\frac{X_1(\tau,\omega)}{X_2(\tau,\omega)}\right]=\omega\delta_{\pi\omega}+\epsilon_{\pi\omega}+2\pi k, \quad k=0,\pm1,\pm2,\cdots \quad (3)$$

where $\theta_{\pi\omega}$ and $\delta_{\pi\omega}$ denote the PDOA and the TDOA between two microphones, respectively, at the time τ and the frequency ω, and $\epsilon_{\pi\omega}$ denote a phase distortion due to the ambient noise.

We focus on classify $\theta_{\pi\omega}$ into one of the signal sources, $S_i(\tau,\omega), i=1,\ldots,N_S$. In order to do so, we need to build a posterior probability density as follows:

$$P(S_i|\theta_{\pi\omega})=\frac{p(\theta_{\pi\omega}\mid S_i)P(S_i;\omega)}{p(\theta_{\pi\omega})}=\frac{p(\theta_{\pi\omega}\mid S_i)P(S_i;\omega)}{\sum_{k=1}^{N_S}p(\theta_{\pi\omega}\mid S_k)P(S_k;\omega)} \quad (4)$$

where S_i denotes class indicator representing the i^{th} source.

3. Proposed method

The proposed method is to build (4) without concerning the 2π ambiguity problem. It is because that the input information is the $\theta_{\pi\omega}$ which ranges from $-\pi$ to π. From (4) let the denominator denote the ω-local PDOA model which addresses an evidence probability of $\theta_{\pi\omega}$. Since von Mises PDF[20] models the distribution of circular data, it is suitable to model the PDOA. We define the likelihood $p(\theta_{\pi\omega}\mid S_i)$ in (4) as

$$p(\theta_{\pi\omega}\mid S_i)=v(\theta_{\pi\omega}\mid\mu_{\omega i},\kappa_{\omega i})=\frac{1}{2\pi I_o(\kappa_{\omega i})}exp(\kappa_{\omega i}\cos(\theta_{\pi\omega}-\mu_{\omega i})) \quad (5)$$

where $v(\theta_{\pi\omega}\mid\mu_{\omega i},\kappa_{\omega i})$ represents the von Mises PDF[20] of $\theta_{\pi\omega}$ with the mean angle $\mu_{\omega i}$ and the concentration $\kappa_{\omega i}$ of the source S_i. The $I_o(\cdot)$ is the modified Bessel function of order 0.

Let z_i denote a latent variable that indicates to which source the parameters of von Mises PDF $\mu_{\omega i}$ and $\kappa_{\omega i}$ belong.

$$z_i=\begin{cases}1, & \text{if}\ \mu_{\omega i},\kappa_{\omega i}\in S_i\\ 0, & otherwise\end{cases} \quad (6)$$

We reform $P(S_i;\omega)$ as

$$P(S_i;\omega)=p(z_i=1;\omega)=\gamma_{\omega i} \quad (7)$$

Using (5) and (7), the ω-local PDOA model, i.e. denominator of (4), is formed as

$$p(\theta_{\pi\omega})=\sum_{i=1}^{N_S}\gamma_{\omega i}v(\theta_{\pi\omega}|\mu_{\omega i},\kappa_{\omega i})\ s.t.\sum_{i}^{N_S}\gamma_{\omega i}=1, \quad (8)$$

and its parameter set $\{\gamma_{\omega i},\mu_{\omega i},\kappa_{\omega i}\}\forall\,\omega,i$ needs to be estimated.

However, the problem is that there may not be sufficient data to estimate those parameters per frequency, ω. For example, if we set stepsize 32 ms for STFT, there are nearly 93 data for 3 seconds of utterance. Considering speech-inactive TF-bin and the signal sparseness, this number can be further reduced. This problem may be an obstacle to implementing a real-time source separator.

To resolve such a data insufficiency problem, we need to utilize all the PDOA data from all frequency range. In order to do so, we need new statistics that can localize signal sources and also can be extracted from all frequencies. As such

statistics, we adopted an incident angle of signal, ϑ, that represents an angle between the source location and the microphone array. The reason why the incident angle was chosen is that each signal source has its own incident angle and its value can be extracted from the TF-bin for all frequencies. Moreover, we can employ the incidence angle distribution as *a priori* and adapt its parameter to build the ω-local PDOA distribution since the conjugate prior of the von Mises distribution is also the von Mises distribution[21].

The ω-local PDOA values are converted to an incident angle using the following equation:

$$\vartheta = \cos^{-1}\left(\frac{c}{\omega d}\theta\right) \tag{9}$$

where d and c are the innerspace length between microphones and the speed of sound

3.1 Incidence angle distribution as *a priori*

The purpose of building the incidence angle distribution is to make a reference model for ω-local PDOA distribution. The distribution of incidence angle ϑ is modeled with the MovM. Since this distribution is irrelevant with the frequency of spectral bin, ω, the entire PDOA data over all frequency bands can be utilized to learn the model parameter for robust estimation. The distribution is formed as

$$p(\vartheta) = \sum_{i=1}^{N_S} \gamma_i v(\vartheta \mid \mu_i, \kappa_i) \;\; s.t. \sum_{i=1}^{N_S} \gamma_i = 1 \tag{10}$$

where γ_i, μ_i and κ_i denote, respectively, the mixture weight, mean angle, and concentration parameter. These parameters were estimated via the Expectation- Maximization (EM) algorithm. We present the resultant formulae here, but readers are encouraged to refer to Banerjee[22] and Calderara[23] for the detail. Let $\vartheta_l, l = 1,...,L$ denote L observations of the incidence angle converted using (9) but we do not know their signal source. The estimation procedure is as follows:

1) Initialize γ_i, μ_i and κ_i with random values
2) Expectation-step: calculate the probability of signal source given observation, $P(S_i|\vartheta_k)$, as follows:

$$P(S_i|\vartheta_l) = \frac{\gamma_i v(\vartheta_l|\mu_i,\kappa_i)}{\sum_{k=1}^{N_S}\gamma_k v(\vartheta_l|\mu_k,\kappa_k)} \tag{11}$$

3) Maximization-step: update γ_i, μ_i and κ_i

$$\gamma_i \leftarrow \frac{1}{L}\sum_{l=1}^{L} P(S_i|\vartheta_l) \tag{12}$$

$$\mu_i \leftarrow \tan^{-1}\left(\frac{\sum_{l=1}^{L} P(S_i|\vartheta_l)\sin\vartheta_l}{\sum_{l=1}^{L} P(S_i|\vartheta_l)\cos\vartheta_l}\right) \tag{13}$$

$$A(\kappa_i) \leftarrow \frac{\sum_{l=1}^{L} P(S_i|\vartheta_l)\cos(\vartheta_l - \mu_i)}{\sum_{l=1}^{L} P(S_i|\vartheta_l)} \tag{14}$$

κ_i is obtained by inversing $A(\kappa_i)$ in terms of κ_i, namely $A^{-1}(\kappa_i)$. However, $A^{-1}(\kappa_i)$ is mathematically intractable so it is approximated[22] as follows

$$\hat{\kappa_i} \approx \frac{2A(\kappa_i) - A^3(\kappa_i)}{1 - A^2(\kappa_i)} \tag{15}$$

4) Iteration: Step 2) and Step 3) are iterated until the value of γ_i, μ_i and κ_i are converged to steady values.

3.2 MAP adaptation with a mixture of von Mises distribution

As we addressed previously, the PDOA distribution cannot help but being modeled over each frequency band, therefore there may be a lack of data to estimate its parameters. Hence our approach is to obtain those parameters by adapting the parameter of the incidence angle distribution to the ω-local PDOA distribution. We employed the Maximum a Posteriori (MAP) adaptation technique which has been used in speaker verification[24,25].

In order to adapt, we need to define a mapping function to convert the incidence angle

unit back to ω-local PDOA unit as follow:

$$f_{\omega}(\vartheta) \equiv \frac{\omega d}{c}\cos(\vartheta) \tag{16}$$

We first determine the probabilistic alignment of the input PDOA into each mixture component of the incidence angle distribution. That is, for mixture i in the distribution, we compute

$$\gamma_{\tau\omega i} = \frac{\gamma_i v(\theta_{\tau\omega}|f_{\omega}(\mu_i),\kappa_{\omega i})}{\sum_{k=1}^{N_S}\gamma_k v(\theta_{\tau\omega}|f_{\omega}(\mu_k),\kappa_{\omega k})} \tag{17}$$

and then compute the sufficient statistics: $\mathsf{S}_{\omega i} = \Sigma_{\tau}\gamma_{\tau\omega j}\sin\theta_{\tau\omega}$ and $\mathsf{C}_{\omega i} = \Sigma_{\tau}\gamma_{\tau\omega j}\cos\theta_{\tau\omega}$ for all ω and i.

We present the resultant estimation formulae here and the detailed derivation can be found in the [21] and [26].

$$\hat{\mu}_{\omega i} = \tan^{-1}\left(\frac{R_i\sin f_{\omega}(\mu_i)+\mathsf{S}_{\omega i}}{R_i\cos f_{\omega}(\mu_i)+\mathsf{C}_{\omega i}}\right) \tag{18}$$

where R_i is the resultant length[20] of data from i^{th} source, which is formed as

$$R_j = W\cdot A(\kappa_j) = W\cdot (I_1(\kappa_j)/I_0(\kappa_j)) \tag{19}$$

while W denotes the number of data used to estimate the parameter of incidence angle distribution and $I_1(\cdot)$ is the modified Bessel function with order 1. The estimate of the concentration parameter, $\kappa_{\omega i}$, is obtained in the same manner of the EM framework by

$$\hat{\kappa}_{\omega i} \approx \frac{2A(\kappa_{\omega i})-A^3(\kappa_{\omega i})}{1-A^2(\kappa_{\omega i})} \tag{20}$$

where

$$A(\kappa_{\omega i}) = \frac{\Sigma_{\tau}\cos(f_{\omega}(\mu_i)-\theta_{\tau\omega})}{W}. \tag{21}$$

Finally, the mixture weight is obtained as follow

$$\hat{\gamma}_{\omega i} = \frac{1}{T}\sum_{\tau=1}^{T}\gamma_{\tau\omega i} \tag{22}$$

3.3 Time-frequency binary masking and signal recovery

In this paper we construct a binary mask using (9) as follow:

$$M_i(\tau,\omega) = \begin{cases} 1 & \text{if } \ i = \underset{k=1,\ldots,N_S}{\arg max} P(S_k|\theta_{\tau\omega}) \\ 0 & otherwixe \end{cases} \tag{23}$$

The separated TF signal for ith source is formed as:

$$\hat{S}_i(\tau,\omega) = M_i(\tau,\omega)X_1(\tau,\omega), i = 1,\ldots,N_S \tag{24}$$

and then $\hat{S}_i(\tau,\omega)$ is transformed to time-domain via the inverse STFT and the overlap-and-add method for all each source.

4. Experiments

To validate the effectiveness of the proposed method, we applied it to the speech separation task and evaluated it in terms of several objective scores. The dataset and measurement are provided from the SiSEC speech separation campaign[27] which is available for the public use of BSS algorithm evaluation.

4.1 Task and settings

The performance of the proposed algorithm was evaluated based on its ability to separate the mixed voices of three speakers at different locations. Those speakers' are either a group of females or males drawn from the 'dev1' set of SiSEC 2008 database[27]. Among the data, we picked up the audio data that was lively recorded in a reverberant environment where the reverberation time was either 130ms or 250ms. The spacing between microphones was set to 5cm or 1m.

4.2 Performance measure

Four objective evaluation scores are measured per data set. Those are Signal to Distortion Ratio (SDR), Source Image to Spatial Distortion Ratio (ISR), Source to Interference Ratio (SIR), and Sources to Artifacts Ratio (SAR). All measures are dB scaled, and larger numbers mean better performance. The detail of those measures can be found in [28].

4.3 Results

In Table 1 and Table 2, we organized the result of the evaluation in terms of condition and the performance measure. In addition, 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 data set. The first method, denoted as "Ozerov"[29], estimates mixing parameters using the expectation-maximization algorithm. The second one, denoted as "Nesta[1]"[30], estimates mixing parameters with the natural gradient algorithm. The third method, denoted as "Nesta[2]" is all the same with [30] but they added an additional Wiener post-processing to recover the original speech signal. The last method, denoted as "Iso"[31], is a combination of the bin-wise clustering method[6] and a full-rank method[8]. Since the evaluation score exhibit per the separated source, we averaged and posted them.

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
5cm	SDR	3.57	3.4	5.2	5.1	**5.9**
	ISR	7.34	9.6	8.7	9.7	**10.4**
	SIR	9.01	8.9	9.2	**10.3**	9.4
	SAR	7.42	7.7	9.1	8.1	**9.9**
1m	SDR	3.52	9.1	7.1	8.2	**8.5**
	ISR	8.83	**14.6**	9.9	13.1	13.3
	SIR	7.58	**14.5**	12.1	13.8	12.9
	SAR	6.68	**11.9**	10.7	10.3	11.7

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
5cm	SDR	2.48	3.7	5.5	**6.2**	4.7
	ISR	3.62	8.9	9.4	**10.5**	8.8
	SIR	5.42	7.2	9.0	**10.5**	7.8
	SAR	8.54	8.0	8.4	8.7	**9.1**
1m	SDR	2.28	7.0	5.6	7.1	**7.5**
	ISR	6.30	**12.3**	8.6	11.5	11.6
	SIR	4.20	**11.7**	9.8	11.6	11.2
	SAR	5.56	9.5	8.4	9.5	**11.1**

The proposed algorithm shows competitive performance in the case of 5 cm microphone spacing and 130ms reverberation time. Note that the proposed method only consists of the spectral-bin clustering and a simple binary mask. Considering that the quality of the separated speech is highly dependent on a series of the post-processing such as mixing parameter estimation, permutation alignment, and signal recovery technologies, it shows the effectiveness of the proposed clustering method. We expect the combination of other post-processing methods with the proposed clustering method can improve performance of other conditions.

Since the proposed algorithm is designed to build appropriate probability distribution for classifying spectral bin, direct comparisons with other methods listed in [28] are meaningless. This is because the proposed algorithm is not meant for speech enhancement or interference suppression. Therefore future work may include mixing parameter estimation over the classified bin for BSS. Instead, we give attention to the SAR since the performance of the binary mask is highly relevant to how much artifacts (a.k.a musical noise) occurred in the separated signal as a result of a wrongly classified spectral bin.

5. Conclusion

Circular statistics based PDOA distribution modeling is proposed for the spectral bin classification. The PDOA distribution is obtained per frequency by using MovM. To resolve data insufficiency problem, we adopt MAP adaptation approach. As an a priori for MAP adaptation, the incidence angle distribution is approximated from the PDOA over the entire frequency band and its parameters are estimated via EM algorithm. In order to do so, statistics of incidence angle is converted to those of PDOA and vice versa. By using a public SiSEC 2008

database, the performance was evaluated in terms of speech enhancement of which task is underdetermined blind source separation. With only the spectral-bin clustering and the simple binary mask, the proposed method shows competitive performance in some environmental conditions. We expect the performance of BSS will be further improved if the proposed method is incorporated with post-processing technologies. The future work includes mixing parameter estimation for sound source separation from the classified spectral bin.

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. 114, no. 4, pp. 1830-1847, 2004. DOI: https://doi.org/10.1109/TASL.2010.2050716

[4] Y. Izumi, N. Ono and S. Sagayama, "Sparseness-based 2ch BSS using the EM algorithm in reverberant environment," in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Palz, 2007.
DOI: http://dx.doi.org/10.1109/ASPAA.2007.4393015

[5] 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.
DOI: http://dx.doi.org/10.1007/978-3-642-00599-2_93

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

[7] 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, pp. 1-12, 2007.
DOI: http://dx.doi.org/10.1155/2007/24717

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

[9] M. I. Mandel, D. P. Ellis and T. Jebara, "An EM algorithm for localizing multiple sound sources in reverberant environments," in NIPS, 2007.
DOI: http://dx.doi.org/10.7551/mitpress/7503.003.0124

[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. 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

[14] 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

[15] 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. pp. 2112-2121, 2014.
DOI: https://doi.org/10.1109/TASLP.2014.2361023

[16] 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. pp. 1-18, 2016.
DOI: http://dx.doi.org/10.1186/s13636-016-0085-x

[17] 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. pp. 1075-1084, 2017.
DOI: https://doi.org/10.1109/TASLP.2017.2687104

[18] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," Signal processing, vol. 81, no. 11, pp. pp. 2403-2418, 2001.
DOI: https://doi.org/10.1016/S0165-1684(01)00128-1

[19] Y. Wang and M. Brookes, "Model-Based Speech Enhancement in the Modulation Domain," IEEE/ACM Trans. on Audio, Speech and Lang. Proc., vol. 26, no. 3, pp. pp. 580-594, 2018.
DOI: https://doi.org/10.1109/TASLP.2017.2786863

[20] K. V. Mardia and P. E. Jupp, Directional Statistics, Wiley, 1999.
DOI: https://doi.org/10.1002/9780470316979

[21] K. V. Mardia and S. A. M. El-Atoum, "Bayesian inference for the von Mises-Fisher distribution," Biometrika , vol. 63, no. 1, pp. 203-206, 1976.
DOI: https://doi.org/10.2307/2335106

[22] A. Banerjee, I. S. Dhillon, J. Chosh and S. Sra, "Clustering on the unit hypersphere using von Mises-Fisher distributions," JMLR, vol. 6, pp. 1345-1382, 2005.

[23] A. P. R. C. S. Calderara, "Mixtures of von Mises Distributions for People Trajectory Shape Analysis," IEEE Trans. Circuits and Systems for Video Technology, vol. 21, no. 4, pp. 457-471, 2011.
DOI: https://doi.org/10.1109/TCSVT.2011.2125550

[24] D. A. Reynolds, T. F. Quatieri and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," Digital Signal Processing, vol. 10, pp. 19-41, 2000.
DOI: https://doi.org/10.1006/dspr.1999.0361

[25] J. -L. Gauvain and C. -H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," IEEE Transactions on Speech and Audio Processing, vol. 2, no. 2, pp. 291-298, 1994.
DOI: https://doi.org/10.1109/89.279278

[26] P. Guttorp and A. Lockhart, "Finding the location of a signal: a Bayesian analysis," Journal of the American Statistical Association, vol. 83, no. 402, pp. 322-330, 1988.
DOI: https://doi.org/10.2307/2288846

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

[28] [Online]. Available: http://www.irisa.fr/metiss/SiSEC11/underdetermined/underdetermined_dev1_mean_speech3_all.html

[29] 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

[30] 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

[31] 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

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