音源到来方向分布とアクティベーション共有型非負値行列因子分解を用いた音像深度推定*

☆宮内 智, 北村大地, 猿渡 洋, 中村 哲 (奈良先端大)

1 Introduction

Wave field synthesis (WFS) [1] is a technique for realizing high quality sound reproduction. WFS allows us to create sound images at the front of loudspeakers. However, WFS requires specific localization information of the objective sound sources in a mixed signal to reproduce the individual sound images. In existing media contents such as a CD, the localization of each sound image has been lost because they are generally provided as a stereo format. Therefore, a method to estimate the localization of each sound image even in the mixed source requires urgent attention. In our previous research, the estimation method of sound directions has been proposed [2]. However, as far as we know, there is no effective method and studies that cope with depth information tracking for sound images. To solve this problem, in this paper, we propose a new depth estimation method, which utilizes the variation of the direction of arrival (DOA) of the individual sound source. In this paper, first, we describe the source separation method based on directional clustering for the mixed source into the individual source. Next, we propose a new depth estimation method, which utilizes the variation of DOA. In addition, we propose the feature extraction method using nonnegative matrix factorization (NMF) [3]. Finally, we show the efficacy of the proposed method by the objective experiment.

2 Source Separation Method

In the proposed method, we use the source separation method based on directional clustering [4] for extracting the specific source included in the mixed signal. Figure 1 shows the configuration of directional clustering. First, the time-frequency components of the stereo mixed signal $\boldsymbol{X}(\omega,\tau) = [X^{(\mathrm{L})}(\omega,\tau), X^{(\mathrm{R})}(\omega,\tau)]^{\mathrm{T}}$ are represented into the two-dimensional space that has $|X^{(\mathrm{L})}(\omega,\tau)|$ and $|X^{(\mathrm{R})}(\omega,\tau)|$ as coordinate axes, where $|X^{(\mathrm{L})}(\omega,\tau)|$ and $|X^{(\mathrm{R})}(\omega,\tau)|$ are the amplitude of each channel. Next, these components are normalized and separated by k-means clustering. Obtained components of the individual cluster are used for reproduction by WFS.

3 Proposed Method

3.1 Depth Estimation Based on DOA Distribution

In sound fields, when a sound source is far from the listener, sound waves arrive from various directions owing to sound diffusion. Therefore, as shown in Fig. 2, the shape of an observed DOA distribution of the target source can be used as a cue for depth estimation. Process flow of the proposed method are shown in Fig. 3. In this method, first, we calculate a weighted DOA histogram of a mixed sound source. In this process, DOAs are calculated as $\theta = 2\arctan\left(|X^{(\mathrm{R})}(\omega,\tau)| \,/\, |X^{(\mathrm{L})}(\omega,\tau)|\right)$. Simultaneously, DOAs are weighted by the magnitude of each vector $w = \left(|X^{(\mathrm{L})}(\omega,\tau)|^2 + |X^{(\mathrm{R})}(\omega,\tau)|^2\right)^{1/2}$ to avoid the problem of normalization process. Next, we separate the mixed source into the individual sources by using directional source separation method. In addition, we propose to extract features of the target components to eliminate the effect of background noise and artificial distortion generated by signal processing (windowing in short-time DFT, etc.). As the feature extraction method, we propose activation-shared multichannel NMF, which reduces dimensionality of input stereo data while maintaining directional information. This property is advantageous to polish up the DOA distribution, whereas the conventional NMFs applied to stereo signals in parallel generate an artificial fluctuation in DOA. Finally, we can estimate the depth of the target source by modeling the resultant DOA distribution. In the following sections, we will describe activation-shared multichannel NMF and the modeling method for DOA distribution.

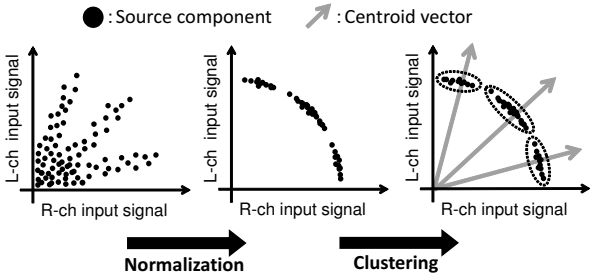

Fig. 1 Configuration of directional clustering.

3.2 Activation-Shared Multichannel NMF

NMF is a sparse representation method. The aim of sparse representations is to reveal certain structures of a signal, and to represent these structures in a compact. Also, this representations provide high performance for noise reduction, compression and feature extraction. Using this property, we eliminate background noise and artificial distortion, which are the problem to evaluate the shape of DOA distribution. However, if the conventional NMFs are applied to stereo signals in paral-

* "Depth estimation of sound images using direction of arrival distribution and activation-shared nonnegative matrix factorization," by Tomo Miyauchi, Daichi Kitamura, Hiroshi Saruwatari, and Satoshi Nakamura (Nara Institute of Science and Technology).

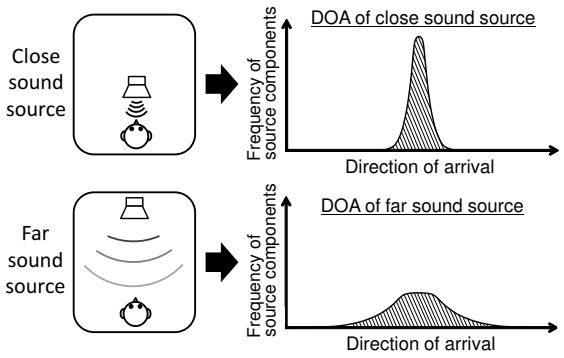

Fig. 2 Example of DOA histogram.

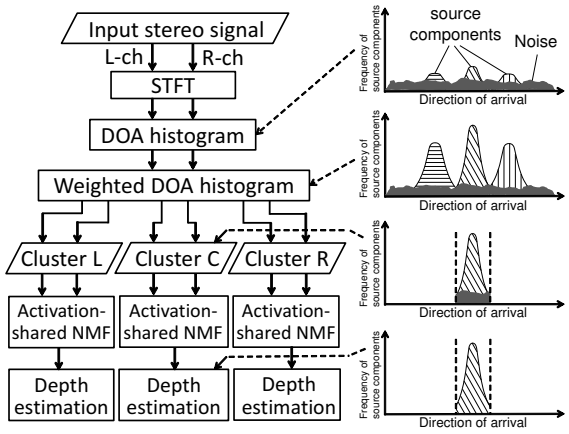

Fig. 3 Signal flow of proposed method.

lel, DOA information is disturbed. Therefore, we propose activation-shared multichannel NMF, which provides sparse representation of the signal while maintaining the directional information. In the following section, we describe the update rules of the proposed NMF, which have been derived by the authors.

3.2.1 Cost Function and Update Rules

When M-channel signals are observed, the trained bases are constructed as

$$\boldsymbol{Y}^{(m)} \simeq \boldsymbol{F}^{(m)}\boldsymbol{G} \quad (m = 1,\ 2,\ \cdots,\ M), \tag{1}$$

where $\boldsymbol{Y}^{(m)}(\in \mathbb{R}^{\Omega\times T}_{\geq 0})$ is a spectrogram of the mth channel signal, $\boldsymbol{F}^{(m)}(\in \mathbb{R}^{\Omega\times K}_{\geq 0})$ is a matrix that involves bases of each channel signal, and $\boldsymbol{G}(\in \mathbb{R}^{K\times T}_{\geq 0})$ is an activation matrix that is shared in all channels of the signals. In adition, Ω is the number of frequency bins, K is the number of bases, and T is the number of frames of the signal. Since the activation matrix is shared through all channels, we can reduce dimensionality of the input signal while maintaining directional information that is amplitude ratio between each channel.

Next, we describe the cost function of proposed NMF. β-divergence is a generalized divergence of variable x corresponding to y [5]. β-divergence is defined as

$$\mathcal{D}_\beta(y\|x) = \begin{cases} \dfrac{y^\beta}{\beta(\beta-1)} + \dfrac{x^\beta}{\beta} - \dfrac{yx^{\beta-1}}{\beta-1} & (\beta \in \mathbb{R}_{\setminus\{0,1\}}) \\ y(\log y - \log x) + x - y & (\beta \to 1) \\ \dfrac{y}{x} - \log\dfrac{y}{x} - 1 & (\beta \to 0) \end{cases}. \tag{2}$$

Using β-divergence, the cost function of activation-shared multichannel NMF is defined as

$$\mathcal{J}(\Theta) = \textstyle\sum_m \mathcal{D}_\beta(\boldsymbol{Y}^{(m)}\|\boldsymbol{F}^{(m)}\boldsymbol{G}), \tag{3}$$

where $\Theta = \left\{\boldsymbol{F}^{(m)}, \boldsymbol{G}\right\}$ is a set of observed variables. Using (2), we can redefine (3) as

$$\mathcal{J}(\Theta) = \textstyle\sum_{m,\omega,t}\left[\dfrac{\left(\sum_k f^{(m)}_{\omega,k} g_{k,t}\right)^\beta}{\beta} - \dfrac{y^{(m)}_{\omega,t}\left(\sum_k f^{(m)}_{\omega,k} g_{k,t}\right)^{\beta-1}}{\beta-1}\right], \tag{4}$$

where $y^{(m)}_{\omega,t}$, $f^{(m)}_{\omega,k}$, and $g_{k,t}$ are entries of matrices $\boldsymbol{Y}^{(m)}$, $\boldsymbol{F}^{(m)}$, and $\boldsymbol{G}$, respectively. Also, the constant term is abbreviated. Then, the update rules are obtained by the auxiliary function approach as

$$f^{(m)}_{\omega,k} \leftarrow f^{(m)}_{\omega,k}\left(\frac{\sum_t y^{(m)}_{\omega,t} g_{k,t}\left(\sum_{k'} f^{(m)}_{\omega,k'} g_{k',t}\right)^{\beta-2}}{\sum_t g_{k,t}\left(\sum_{k'} f^{(m)}_{\omega,k'} g_{k',t}\right)^{\beta-1}}\right)^{\varphi(\beta)}, \tag{5}$$

$$g_{k,t} \leftarrow g_{k,t}\left(\frac{\sum_{m,\omega} y^{(m)}_{\omega,t} f^{(m)}_{\omega,k}\left(\sum_{m,k'} f^{(m)}_{\omega,k'} g_{k',t}\right)^{\beta-2}}{\sum_{m,\omega} f^{(m)}_{\omega,k}\left(\sum_{m,k'} f^{(m)}_{\omega,k'} g_{k',t}\right)^{\beta-1}}\right)^{\varphi(\beta)}, \tag{6}$$

where $\varphi(\beta)$ is defined as

$$\varphi(\beta) = \begin{cases} 1/(2-\beta) & (\beta < 1) \\ 1 & (1 \leq \beta \leq 2) \\ 1/(\beta-1) & (2 < \beta) \end{cases}. \tag{7}$$

3.3 Modeling of DOA distribution

There are many studies on the modeling of DOA distribution, and the modeling methods using a Cauchy distribution and a Gaussian distribution have been proposed in conventional works [6]. However, since these distributions do not have parameters that handle a tail of the distribution accurately, it is not suitable for the "depth" estimation. Therefore, in our method, we evaluate the shape of DOA by using generalized Gaussian distribution (GGD) [7]. GGD is a flexible family of probability density function (PDF) modeling with some parameters, and GGD can represent various types of well-known PDFs, e.g., Gaussian and Laplacian distributions. The definition of GGD is

$$f_{GG}(z;\alpha_{\text{scale}},\beta_{\text{shape}}) = \frac{\beta_{\text{shape}}}{2\alpha_{\text{scale}}\Gamma(\frac{1}{\beta_{\text{shape}}})}\exp\left(-\left[\frac{|z-\bar{z}|}{\alpha_{\text{scale}}}\right]^{\beta_{\text{shape}}}\right), \tag{8}$$

where $\Gamma(x) = \int_0^\infty \exp(-t)t^{x-1}\mathrm{d}t$ is gamma function, $\bar{z}$ is the mean of variable z, α_{scale} is *scale parameter*, β_{shape}

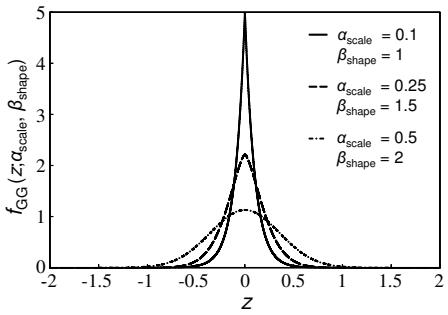

Fig. 4 GGD with typical parameters α_{scale} and β_{shape}.

is *shape parameter* of GGD. Figure 4 shows examples of different PDFs in GGD. As can be seen, the shape of GGD changes depending on β_{shape} ; note that $\beta_{\text{shape}} = 2$ corresponds to Gaussian PDF and that $\beta_{\text{shape}} = 1$ corresponds to Laplacian PDF. If β_{shape} is low, GGD becomes a spiky shape, and if β_{shape} is high, GGD becomes a smooth shape. Based on this property, we use the shape parameter β_{shape} of GGD to evaluate the shape of DOA distribution. In our depth estimation method, depth of sound images are evaluated by β_{shape} based on the definition that the target source is close when β_{shape} is low, and the target source is far when β_{shape} is high.

Although the maximum-likelihood based shape parameter estimation has no closed-form solution in GGD, we propose a closed-form parameter estimation algorithm based on some approximation and kurtosis. We will describe this estimation method as below.

3.3.1 Closed-Form Parameter Estimation in GGD

We can introduce *kurtosis* to estimate β_{shape}. In general, kurtosis of DOA histogram $h(\theta)$ is given by

$$\text{kurt}(h(\theta)) = \langle h^4(\theta)\rangle_\theta \langle h^2(\theta)\rangle_\theta^{-2} - 3. \tag{9}$$

The nth order moment of GGD has the following useful relationship;

$$\langle z^n\rangle = \beta_{\text{shape}}^{-n}\Gamma\left(\frac{n+1}{\beta_{\text{shape}}}\right)\Gamma\left(\frac{1}{\beta_{\text{shape}}}\right)^{-1}. \tag{10}$$

From (9) and (10) we have the following equation of kurtosis and β,

$$\text{kurt}(y(\theta)) = \Gamma\left(\frac{5}{\beta_{\text{shape}}}\right)\Gamma\left(\frac{1}{\beta_{\text{shape}}}\right)\Gamma\left(\frac{3}{\beta_{\text{shape}}}\right)^{-2} - 3. \tag{11}$$

In order to obtain β_{shape} value from the measured kurtosis, we should calculate the inverse function of (11). However, it is well known that there is no exact closed-form solution of (11) w.r.t. β_{shape}. Therefore, we hereinafter introduce an approximation for the closed-form derivation of optimal β_{shape}. With modified Stirling's formula on gamma function,

$$\Gamma(z) \sim \sqrt{2\pi}\cdot\exp(-z)\cdot z^{z-0.5}\cdot\exp\left(\frac{1}{12z}\right) \tag{12}$$

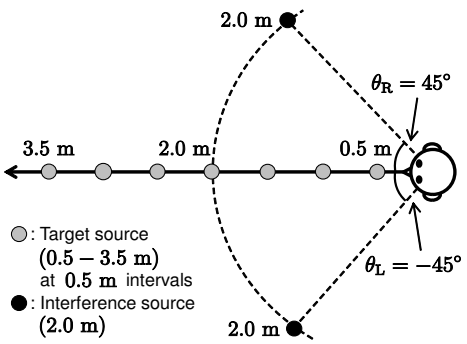

Fig. 5 Sound source geometry used in experiment.

or equally

$$\log\Gamma(z) \sim \frac{1}{2}\log(2\pi) - z + (z-0.5)\log z + \frac{1}{12z}, \tag{13}$$

then we can obtain

$$\begin{aligned}&\log(\text{kurt}(h(\theta)) + 3)\\ &= \log\Gamma\left(\frac{5}{\beta_{\text{shape}}}\right) + \log\Gamma\left(\frac{1}{\beta_{\text{shape}}}\right) - 2\log\Gamma\left(\frac{3}{\beta_{\text{shape}}}\right)\\ &\sim \frac{1}{\beta_{\text{shape}}}\cdot\log\left(\frac{5^5}{3^6}\right) + \log\left(\frac{3}{\sqrt{5}}\right) + \beta\cdot\frac{2}{45}.\end{aligned} \tag{14}$$

This results in the following quadratic equation of β_{shape} to be solved,

$$\beta_{\text{shape}}^2 + \beta_{\text{shape}}\cdot\frac{45}{2}\log\left(\frac{3}{\sqrt{5}(\text{kurt}(h(\theta))+3)}\right) + \frac{45}{2}\log\left(\frac{5^5}{3^6}\right) = 0, \tag{15}$$

and we can derive the closed-form estimate of shape parameter β_{shape} from kurtosis, as

$$\begin{aligned}\beta_{\text{shape}} = &-\frac{45}{4}\log\left(\frac{3}{\sqrt{5}(\text{kurt}(h(\theta))+3)}\right)\\ &-\frac{1}{2}\sqrt{\frac{2025}{4}\log\left(\frac{3}{\sqrt{5}(\text{kurt}(h(\theta))+3)}\right)^2 - 90\log\left(\frac{5^5}{3^6}\right)}.\end{aligned} \tag{16}$$

4 Experiment and Results

A geometry of an experiment is shown in Fig. 5. We prepared two mixed stereo signals containing three instruments. Mixed signal 1 consists of vocal (Vo.), piano (Pf.), and guitar (Gt.). Also, mixed source 2 consists of Vo., electric guitar (E.Gt.), and synthesizer (Syn.). The target source was located in the center direction with seven distances. In addition, the interference sources were located in the left- and right-hand sides. Test sources were generated using a room impulse response recorded at each position. We used the image method [8] as a reference for this experiment, which is a technique of simulating the room impulse response. The signal not processed by NMF was evaluated as conventional

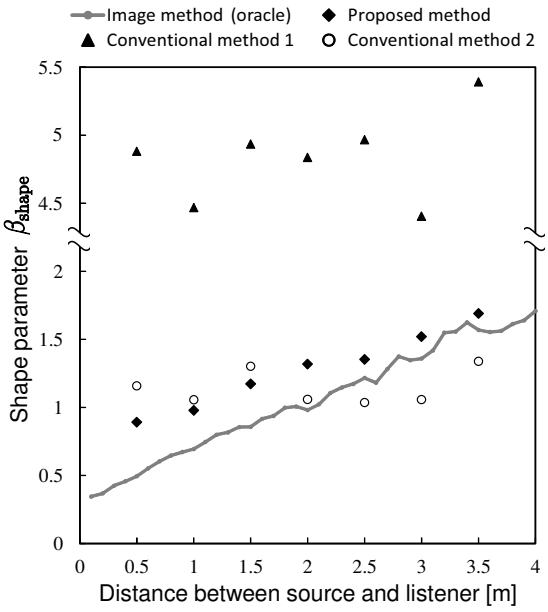

Fig. 6 Estimated value of mixed signal 1.

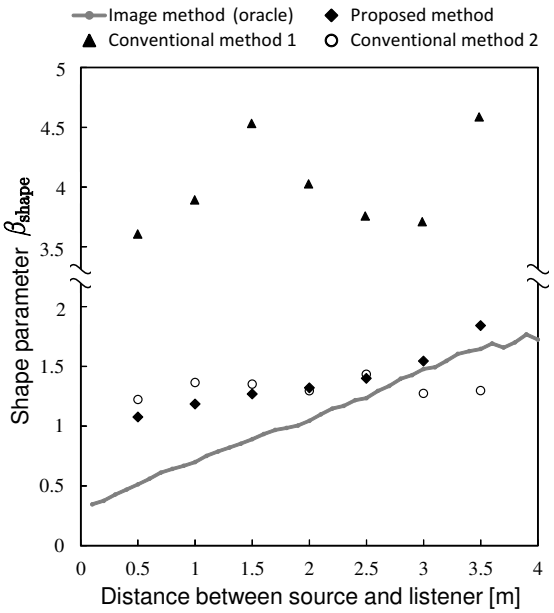

Fig. 7 Estimated value of mixed signal 2.

Table 1 Correlation coefficient of mixed signal 1

Data set	1	2	3	4	5	6
Target source	Vo.	Vo.	Gt.	Gt.	Pf.	Pf.
Interference source (left)	Pf.	Gt.	Pf.	Vo.	Vo.	Gt.
Interference source (right)	Gt.	Pf.	Vo.	Pf.	Gt.	Vo.
Proposed method	0.903	0.891	0.777	0.651	0.791	0.816
Conventional method 1	0.419	0.532	0.154	0.277	0.602	0.496
Conventional method 2	0.233	0.165	0.044	-0.037	0.426	0.157

Table 2 Correlation coefficient of mixed signal 2

Data set	1	2	3	4	5	6
Target source	Vo.	Vo.	E.Gt.	E.Gt.	Syn.	Syn.
Interference source (left)	Syn.	E.Gt.	Syn.	Vo.	Vo.	E.Gt.
Interference source (right)	E.Gt.	Syn.	Vo.	Syn.	E.Gt.	Vo.
Proposed method	0.901	0.882	0.797	0.899	0.940	0.735
Conventional method 1	0.353	0.298	0.127	0.419	0.419	0.042
Conventional method 2	0.116	0.208	-0.049	0.457	0.239	0.370

method 1. Also, the signal processed by conventional NMF, which was applied to each channel independently, was evaluated as conventional method 2.

The experimental results are shown in Fig. 6 and Fig. 7. From these results, the shape parameters of the proposed method are proportional to the distance of the target source, whereas the results of the conventional methods have no agreement with the oracle. In addition, the correlation coefficient between the reference value of image method and the estimated value of other methods are shown in Table 1 and Table 2. These results indicate strong relation between the estimated value of the proposed method and the distance of the target source. Thus, the efficacy of the proposed method as the depth estimation can be confirmed.

5 Conclusion

In this paper, we proposed a new depth estimation method of sound source in mixed signal using the DOA distribution. We also proposed a new feature extraction method for the multichannel signal, activation-shared NMF. The result of the experiment indicated the efficacy of the proposed method.

References

[1] A. J. Berkhout, “A holographic approach to acoustic control,” *J. Audio Eng. Soc.*, vol.36, no.12, pp.977–995, 1988.

[2] M. Hirata, et al., “Spatial depth perception of focusing source by wave field synthesis for 3D sound reproduction," *Proc. 3DSA*, pp.137–140, 2011.

[3] D. D. Lee, H. S. Seung, “Algorithms for non-negative matrix factorization," *Neural Info. Process. Syst.*, vol.13, pp.556–562, 2001.

[4] S. Araki, H. Sawada, R. Mukai, S. Makino, “Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors,” *Signal Processing*, vol.87, no.8, pp.1833–1847, 2007.

[5] S. Eguchi, Y. Kano, “Robustifying maximum likefood estimation," *Technical Report of Institute of Statistical Mathematics*, 2001.

[6] M. Hirakawa, K. Suyama, “Multiple sound source tracking by two microphones using PSO," *Proc. ISPACS*, pp.467–479, 2013.

[7] G. Box, et al., “Bayesian Inference in Statistical Analysis,” Addison Wesley, Reading, Massachusetts, 1973.

[8] B. Allen and A. Berkley “Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.,* vol.65, no.4, pp.943–950, 1979.