# Effective Sound Source Separation Using Single Voice Activity Segments for Binaural Sound

Wataru NOGUCHI* and Arata KAWAMURA† and Youji IIGUNI*
* Osaka University, Osaka, Japan
E-mail: noguchi@sip.sys.es.osaka-u.ac.jp Tel/Fax: +06-6850-6580
† Kyoto Sangyo University, Kyoto, Japan
E-mail:kawamura@cc.kyoto-su.ac.jp Tel/Fax: +075-705-1856

***Abstract*—Binaural sound is well known as a three dimensional sound which achieves a virtual reality of sound, or an augmented reality of sound. Binaural sound is characterized by ILD (Interaural Level Difference) and IPD (Interaural Phase Difference). In this paper, we propose a method to separate sound sources from an observed mixture signal, without losing respective ILD and IPD. The proposed method is established by improving a conventional sound source separation method based on single voice activity detection which detects segments including only single sound source. The proposed method estimates ILD and IPD from the single voice activity segments. After separating the sound sources by using the conventional method, we give the estimated ILD and IPD on the separated sound sources to refine the binaural characteristics. The effectiveness of the proposed method is clarified from estimation results of ILD and IPD for binaural sounds.**

## I. Introduction

Binaural sound is a so-called 3D sound, and it reproduces a real sound environment. Binaural sound consists of the left and right sound signals that are corresponding to sound signals observed at the left and right ears. Binaural sound is characterized by ILD (Interaural Level Difference) and IPD (Interaural Phase Difference)[1].

Sound source separation for binaural sound can be applied to sound localization systems[2], 3D audio systems[3], and so on. Many sound source separation methods have been proposed. Adaptive microphone array methods[4] and independent component analysis methods[5] basically require many microphones, i.e., the number of microphones is greater than or equal to the number of sound sources. On the other hand, BSS-BWC (Blind Source Separation via Bin-Wise Clustering)[6] and BSS-SVAD (BSS based on Single Voice Activity Detection)[7] are recently proposed and they require only two microphones. BSS-SVAD is a simple version of BSS-BWC, and achieves low computation load. BSS-SVAD can also be applied to the sound source separation for the binaural sound. When the observed binaural sound includes multiple sound sources they share some frequencies at the same time, ILD and IPD are overlapped. In this case, the separated sound sources by using BSS-SVAD do not recover original ILD and IPD.

In this paper, we investigate a sound source separation method recovering the original ILD and IPD. We extend BSS-SVAD for binaural sounds, by introducing the estimation of ILD and IPD. We estimate ILD and IPD in SV (Single Voice) segments which only single sound source exists. In SV segments, non-overlapped ILD and IPD are easily obtained when the spectral power is sufficiently large. To obtain ILD and IPD with high accuracy, we estimate them only from SV segments with high spectral power. In the proposed method, one separated signal obtained from BSS-SVAD is used to create the other separated sound so that ILD and IPD relation hold. As a result, it is possible to obtain a two-channel separated signal keeping the feature of binaural sound. At the end of this paper, we confirm the effectiveness of the proposed method via simulation. The simulation results showed that the proposed method gives small estimation error for ILD and IPD in comparison to the conventional method.

## II. Blind Source Separation Based on Single Voice Activity Detection

In this section, we give an overview of the sound source separation procedure of BSS-SVAD[7]. The proposed method is obtained by expanding BSS-SVAD.

### A. Sound Source Separation Procedure

Let $x_i(t)$ be the observed signal at the $i$th microphone at time $t$, where $i = 1, 2$. We assume that $x_i(t)$ is given as

$$x_i(t) = \sum_{k=1}^{N} v_{ki}(t) + n_i(t), \tag{1}$$

$$v_{ki}(t) = \sum_{l=0}^{L-1} h_{ki}(l) s_k(t-l), \tag{2}$$

where $N$ is the number of sound sources, $s_k(t)$ denotes the $k$th sound source signal, $n_i(t)$ denotes an environmental noise observed at the $i$th microphone, $h_{ki}$ denotes the impulse response from $s_k(t)$ to the $i$th microphone, and $L$ is the length of the impulse response. The signal $v_{ki}(t)$ denotes the $k$th sound source observed at the $i$th microphone.

The overview of BSS-SVAD is shown in Fig.1. First, taking STFT (Short-Time Fourier Transform) of $x_i(t)$, we have the observed spectrum $X_i(\tau)$ $(i = 1, 2)$ where $\tau$ denotes the frame index. The observed spectrum $X_i(\tau)$ is given as

$$X_i(\tau) = H_{ki} S_k(\tau) + N_i(\tau) + \sum_{k' \neq k} H_{k'i} S_{k'}(\tau), \tag{3}$$

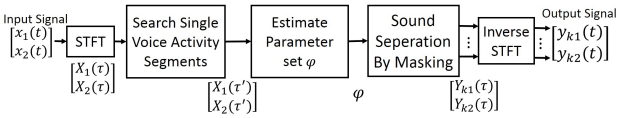



Fig. 1. Overview of BSS-SVAD

where $H_{ki}$, $S_k(\tau)$, $N_i(\tau)$ denote STFTs of $h_{ki}(t)$, $s_k(t)$, $n_i(t)$, respectively.

In the literature[7], the probability density function of $X_2(\tau)/X_1(\tau)$ is modelled by GMM (Gaussian Mixture Model). To get GMM, we estimates its parameters from SV segments. Based on the GMM, we calculate the posteriori probability of $X_i(\tau)$ at each T-F (Time-Frequency) bin. The posteriori probability directly gives a T-F mask which takes a value less than or equal to unit at each T-F bin. The separated spectrum $Y_{ki}(\tau)$ is obtained by multiplying $X_i(\tau)$ with T-F mask. Taking inverse STFT of $Y_{ki}(\tau)$, we have the separated signal $y_{ki}(t)$ in time domain.

Detail explanations of each block shown in Fig.1 are presented in the following subsections.

### *B. Single Voice Activity Detection*

Firstly, we expain how to detect SV segments. The ratio of $X_1(\tau)$ and $X_2(\tau)$ is given as

$$R_X(\tau) = \frac{X_2(\tau)}{X_1(\tau)} = \frac{H_{k2}S_k(\tau) + N_2(\tau) + o_2}{H_{k1}S_k(\tau) + N_1(\tau) + o_1}, \tag{4}$$

$$o_1 = \sum_{k' \neq k} H_{k'1}S_{k'}(\tau), \tag{5}$$

$$o_2 = \sum_{k' \neq k} H_{k'2}S_{k'}(\tau), \tag{6}$$

where $o_i$ denotes the observed spectrum excluding the $k$th sound source at the $i$th microphone. Let $\tau'$ be the frame index of SV segment. We have

$$R_X(\tau') = \frac{H_{k2}S_k(\tau') + N_2(\tau')}{H_{k1}S_k(\tau') + N_1(\tau')}. \tag{7}$$

Here, we define $Q_k$ as

$$Q_k = \frac{H_{k2}}{H_{k1}}. \tag{8}$$

When $|N_1(\tau)|$ and $|N_2(\tau)|$ are sufficiently small in comparison to $|H_{ki}S_k(\tau)|$, we have $R_X(\tau') \approx Q_k$. On the other hand, when $\tau \neq \tau'$, $R_X(\tau)$ fluctuates. Hence, we can judge the present frame $\tau$ as a SV segment when $R_X(\tau)$ does not change in successive past several frames.

### *C. Modeling By GMM*

Next, we describe a method of modeling observed mixture signal with GMM. Let the observed vector be $\boldsymbol{X}(\tau) = [X_1(\tau), X_2(\tau)]^{\mathrm{T}}$. In order to cancel the influence of the amplitude characteristic of the sound source $S_k(\tau)$, $\boldsymbol{X}(\tau)$ is normalized to create a new vector $\boldsymbol{X}'(\tau)$. When a histogram of $\boldsymbol{X}'(\tau)$ is created, $N$ peaks equal to the number of sound sources appear. The histogram is approximated by using GMM. Each cluster is represented by the complex Gaussian density function as

$$p(\boldsymbol{X}'(\tau)|\boldsymbol{a}_k, \sigma_k) = \frac{1}{\pi\sigma_k^2} \exp\left(-\frac{||\boldsymbol{X}'(\tau) - \boldsymbol{a}_k^H \boldsymbol{X}'(\tau)\boldsymbol{a}_k||^2}{\sigma_k^2}\right), \tag{9}$$

where $\boldsymbol{a}_k = [a_{k1}, a_{k2}]^{\mathrm{T}}$ is the average of the Gaussian distribution approximating the $k$th cluster, and $\sigma_k^2$ represents the variance of the Gaussian distribution. By approximating the histogram using (9), we have the following probability density function.

$$p(\boldsymbol{X}'(\tau)|\varphi) = \sum_{k=1}^{N} \beta_k p(\boldsymbol{X}'(\tau)|\boldsymbol{a}_k, \sigma_k), \tag{10}$$

$$\varphi = \{\boldsymbol{a}_1, \sigma_1^2, \beta_1, ..., \boldsymbol{a}_N, \sigma_N^2, \beta_N\}, \tag{11}$$

where $\varphi$ represents the parameter set and $\beta_k$ represents the weight of each Gaussian function. In order to approximate $\boldsymbol{X}'(\tau)$ histogram, it is necessary to estimate a parameter set $\varphi$. BSS-SVAD simply puts $\sigma_k^2$ and $\beta_k$ as

$$\sigma_k^2 = 0.1, \tag{12}$$

$$\beta_k = 1/N. \tag{13}$$

The average $\boldsymbol{a}_k$ is estimated by using Leader-Follower-Clustering (L-F-C)[8] for the phase spectrum $\angle R_X(\tau)$ of $R_X(\tau)$. Here, $C_k$ is a cluster of the $k$th sound source, and $\tau_k'$ denotes a frame of SV segment of $k$th sound source. The center of $C_k$ is estimated as the median value of all $\angle R_X(\tau_k')$ for each frame. We estimate $\boldsymbol{a}_k$ as

$$|\boldsymbol{a}_k| = \mathrm{mean}\{|\boldsymbol{X}'(\tau_{\mathrm{k}}')|\}, \tag{14}$$

$$\angle\boldsymbol{a}_k = \mathrm{mean}\{\angle\boldsymbol{X}'(\tau_{\mathrm{k}}')\}, \tag{15}$$

where, $|\cdot|$ denotes the amplitude spectrum, $\angle\{\cdot\}$ denotes the phase spectrum, and $\mathrm{mean}\{\cdot\}$ denotes the operator that calculates the average value. Determining the parameter set $\varphi$, we can calculate the posteriori probability (10).

### *D. Mask Creation Based on Posteriori Probability*

Based on the posteriori probability $P(C_k|\boldsymbol{X}'(\tau))$ of the sound source $k$, the T-F mask $F_k(\tau)$ is given as

$$F_k(\tau) = P(C_k|\boldsymbol{X}'(\tau)). \tag{16}$$

The spectrum of the separated signal of sound source $k$ is obtained as

$$\boldsymbol{Y}_k(\tau) = F_k(\tau)\boldsymbol{X}(\tau), \tag{17}$$

where $\boldsymbol{Y}_k(\tau) = [Y_{k1}, Y_{k2}]^{\mathrm{T}}$. Taking inverse STFT of $\boldsymbol{Y}_k(\tau)$, a separated signal $y_{ki}(t)$ in time domain is obtained.

## III. Binaural Sound Source Separation

In this section, we describe problems of conventional sound source separation method in binaural sound and give the solution method.

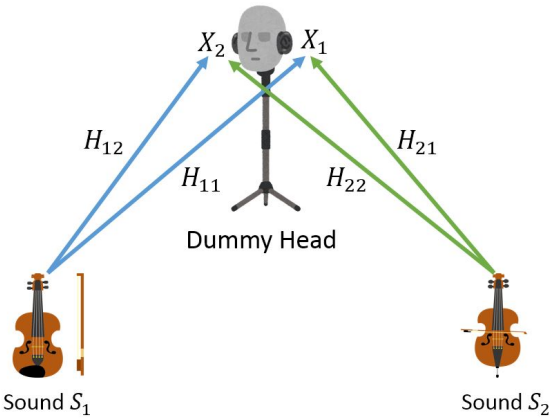



Fig. 2. Image of signal observation using dummy-head

### A. Binaural Sound and Head Related Transfer Function

In binaural sound, it is necessary to reproduce the acoustic effect of the human head. When recording a binaural sound, a dummy head, which is a stereo microphone simulating the human head, is used. Microphones of dummy head are placed at a position corresponding to the human eardrum. It is possible to record a signal in consideration of the HRTF (Head-Related Transfer Function) of the human head. It is said that HRTF plays an important role especially with regard to direction[1]. In this paper, we discuss the perception in the horizontal direction.

Binaural sound can be simulated by convolving the impulse response, recorded by the left and right ears of the dummy head, with an acoustic signal. The image of sound observation by the dummy head is shown in the Fig.2. In Fig.2, $S_1$ and $S_2$ represent sound sources, $X_1$ and $X_2$ represent observed signals, and $H_{k1}$ and $H_{k2}$ $(k = 1,\ 2)$ denote transfer function from sound source to dummy head.

### B. Refine the Binaural Characteristics Based on ILD and IPD

Considering a method that enables binaural characteristic reproduction using ILD and IPD even when left and right HRTFs are unknown. Since both ears of human are attached to both sides of the head, when sound comes in from the side, it arises a difference in sound level and arrival time to both ears. It is known that the amplitude ratio of the binaural sounds is a clue to perception in the left and right direction over the whole audible frequency range. On the other hand, it is said that the time difference (phase difference) from the sound source to the ears is limited to about 1.6kHz or less, which is a clue to the perception of the left and right direction[1]. ILD and IPD at the sound source $k$ with respect to channel-1, which is observed 1st microphone, can be expressed as

$$D_{Lk} = 20\log_{10}(|H_{k2}|) - 20\log_{10}(|H_{k1}|), \quad (18)$$
$$D_{Pk} = \angle H_{k2} - \angle H_{k1}, \quad (19)$$

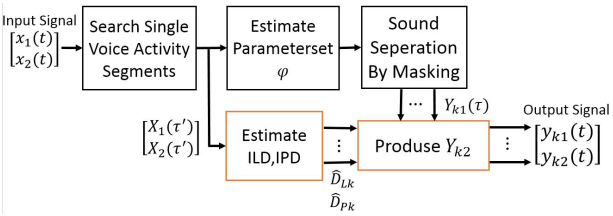



Fig. 3. Flow of proposed method

respectively.Even when HRTF is unknown and $D_{Lk}$ and $D_{Pk}$ are known, spectrum $Y_{k2}$ can be generated from $Y_{k1}$ as

$$Y_{k2} = |Y_{k1}| \cdot 10^{\left(\frac{D_{Lk}}{20}\right)} \exp\{j(\angle Y_{k1} + D_{Pk})\}. \quad (20)$$

### C. Estimate ILD and IPD from Single Voice Activity Segments

In this section, we will derive a sound source separation method effective for binaural sound by using BSS-SVAD.

If $D_{Lk}$, $D_{Pk}$ and the separated signal spectrum of channel-1 are found, the separated signal spectrum of channel-2 which is effective for binaural sound can be obtained from (20).

Since a SV segment is only single sound source speaks, other sound sources do not share the same frequency at the same time. When $D_{Lk}$ and $D_{Pk}$ of each sound source can be estimated from a SV segment of each sound source. In this paper, it is necessary to estimate $D_{Lk}$ and $D_{Pk}$ with high accuracy. The influence of environmental noise etc., the ratio of the spectrum with small power is less reliable than the ratio of spectrum with high power. If the entire spectrum of $\tau'$ is used, the accuracy of estimation of $D_{Lk}$ and $D_{Pk}$ is reduced due to the influence of error. So, we use only the spectrum with high power among $\tau'$.

First, we collect all spectrum of $\tau'$ of sound source $k$. Arrange in descending order of power for each frequency. Then, the spectrum of the upper $T\%$ is extracted for each frequency. We calculate $R_X$ only from each extracted spectrum. And estimate values $\hat{D}_{Lk}, \hat{D}_{Pk}$ of $D_{Lk}, D_{Pk}$ from the average value of amplitude and phase of $R_X$ as

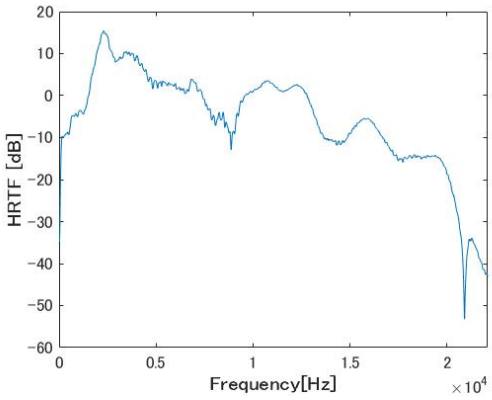
$$\hat{D}_{Lk} = \mathrm{mean}\{20\log_{10}(|\mathrm{R_X}(\tau'_\mathrm{k})|)\}, \quad (21)$$
$$\hat{D}_{Pk} = \mathrm{mean}\{\angle \mathrm{R_X}(\tau'_\mathrm{k})\}, \quad (22)$$

respectively.

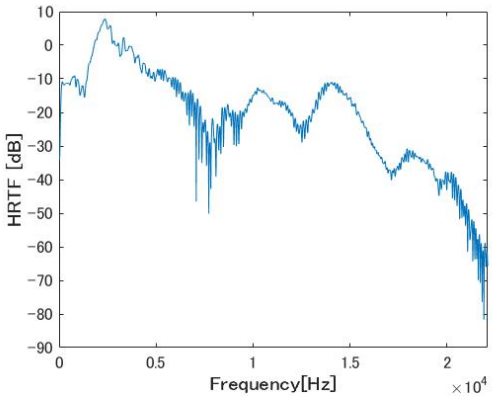
Next, using $\hat{D}_{Lk}$, $\hat{D}_{Lk}$ and the separated signal spectrum $Y_{k1}$, we create separated signal spectrum $Y_{k2}$ from (20).

Block diagram of propose method is shown in Fig.3. First, SV segment is detected from $X_1(\tau)$, $X_2(\tau)$. Second, we estimate $\varphi$, $\hat{D}_{Lk}$, $\hat{D}_{Lk}$ and create a cluster $C_k$. Third, we calculate $P(C_k|X_1(\tau))$. Then, T-F mask is created based on $P(C_k|X_1(\tau))$, and a separated spectrum $Y_{k1}(\tau)$ is obtained. Finally, $Y_{k2}(\tau)$ is created by (20), and taking inverse STFT of $Y_{k1}(\tau)$ and $Y_{k2}(\tau)$ to obtain $y_{k1}(t)$ and $y_{k2}(t)$.

In the proposed method, the features of binaural sound can be retained due to the estimated ILD and IPD.




(a) Left




(b) Right

Fig. 4. HRTF of sound source position is 40°

## IV. Performance Evaluation

We carried out simulations to confirm the capability of the proposed method.

### A. Experimental Conditions

Binaural sounds used in the simulation are created using HRTF[9] distributed from MIT Media Laboratory as experimental data. The HRTF was obtained from a dummy head, where the sound source is located at a distance of 1.4m from the dummy head in an anechoic room and measured sound sampled at 44.1kHz. The HRTF[9] used in the simulation is shown in Fig.4(a) and Fig.**??**, where the sound source is located at $40°$ position to the right from the center of the dummy head. We used an instrumental sound taken from MedleyDB[10] which is music database for research. This sound is a music signal of 10 seconds. The sound source position assumed that in the horizontal plane from the center of the dummy head microphone at $40°$ positions to the right $(R40°)$ and $30°$ positions to the left $(L30°)$. Two types of binaural sounds were created by convolving HRTF of that condition with sound.

Experimental conditions are summarized in a TableI.

TABLE I
Conditions in simulation

| | |
|---|---|
| Number of Source | 2 |
| Direction of Source | $R40°$, $L30°$ |
| Length of Sound | $10[s]$ |
| Sampling Frequency | 44.1kHz |
| Frame Length | 4096 (92ms) |
| Frame Sift Length | 1024 (23ms) |
| $T$ | 10% |

TABLE II
Results of objective evaluation

| Method | Direction | $SD$ | $MSE$ |
|---|---|---|---|
| BSS-SVAD[7] | $R40°$ | 15.7dB | 1.6 |
| Prop. | $R40°$ | 4.1dB | 1.2 |
| BSS-SVAD[7] | $L30°$ | 13.2dB | 1.0 |
| Prop. | $L30°$ | 5.7dB | 0.8 |

### B. Objective Evaluation

In this subsection, we compare the estimation results of ILD and IPD by the conventional method (BSS-SVAD[7]) and the proposed method (Prop.). As an objective evaluation of ILD, We used SD (Spectrum Distortion) which evaluates the difference in amplitude spectrum. SD[1] can be written as

$$SD = \sqrt{\frac{1}{N}\sum_{m=1}^{N}\left[20\log_{10}\frac{|\hat{D}_{Lk}(m)|}{|D_{Lk}(m)|}\right]^2}, \qquad (23)$$

where $\hat{D}_{Lk}(m)$, $D_{Lk}(m)$ denote the estimated and true ILDs, respectively. And $m$ represents frequency index. For good estimation, SD approaches 0.

On the other hand, MSE (Mean Square Error) is used as an objective evaluation for IPD. MSE is defined as

$$MSE_k = \frac{1}{N}\sum_{m=1}^{N}\left(\hat{D}_{Pk}(m) - D_{Pk}(m)\right)^2, \qquad (24)$$

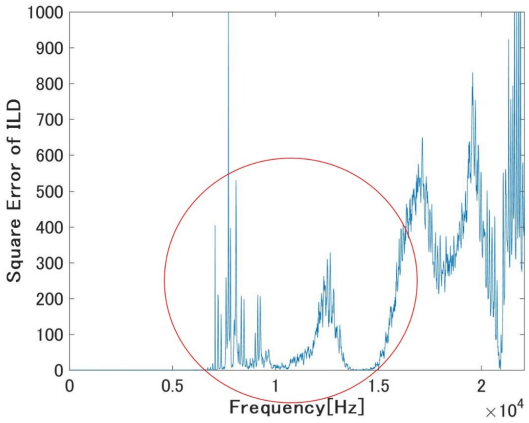
where $\hat{D}_{Pk}(m)$, $D_{Pk}(m)$ represent the estimated and the true IPDs, respectively. For good estimation, MSE approaches 0.

The objective evaluation results are summarized in a TableII. From TableII, SD at $R40°$ was 4.1dB for Prop., and 15.7dB for the BSS-SVAD. $MSE$ of Prop. was 1.2, and BSS-SVAD was 1.6. $SD$ is improved by 11.6dB, and $MSE$ is improved by 0.4 points. Similarly, when the sound source position is $L30°$, $SD$ was 5.7dB for Prop., while BSS-SVAD gave 13.2 dB. MSE of Prop. was 0.8, BSS-SVAD was 1.0. $SD$ is improved by 9.1dB and $MSE$ is improved 0.2 points. From these results, it is found that Prop. can estimate more accurate ILD and IPD than BSS-SVAD.

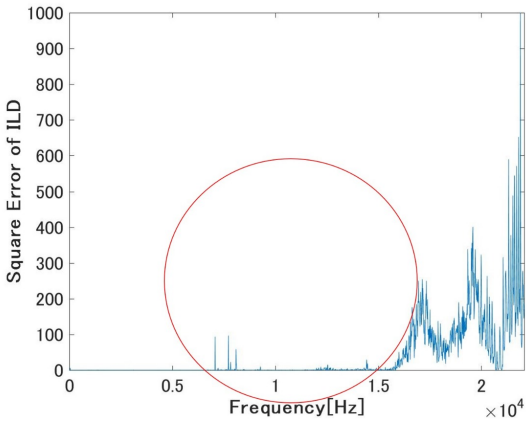
Also, the square error between the estimated value and the true value of ILD and IPD for each frequency is used. The square error $SE_L(m)$, $SE_P(m)$ between the estimated value of ILD and IPD and the true value was calculated as

$$SE_L(m) = \left(\hat{D}_L(m) - D_L(m)\right)^2, \qquad (25)$$

$$SE_P(m) = \left(\hat{D}_P(m) - D_P(m)\right)^2. \qquad (26)$$

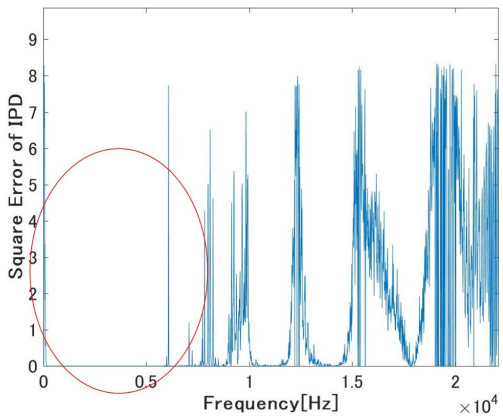



(a) BSS-SVAD[7]

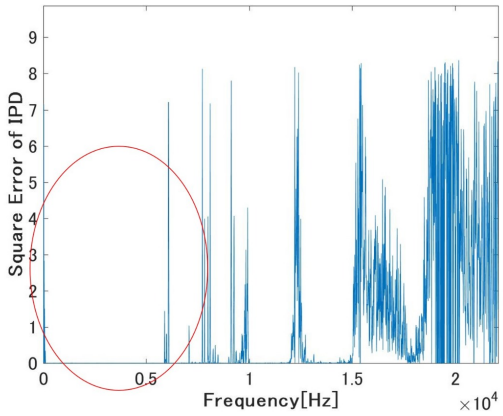



(b) Prop.

Fig. 5. Square error of ILD which sound source position is $R40°$




(a) BSS-SVAD[7]




(b) Prop.

Fig. 6. Square error of IPD which sound source position is $R40°$

Evaluation results were shown in Fig.5(a) to Fig.6(b), where the vertical axis is the square error and the horizontal axis is the frequency.

## V. Conclusions

Comparing Fig.5(a) with 5(b), it can be confirmed that Prop. can estimate the ILD with high accuracy around at about 7kHz to 16kHz. Although the square error is smaller than BSS-SVAD in the frequency higher than 16kHz, the estimation error is large. This reason may be that the sound source used in the simulation did not have sufficiently large power in the frequencies higher than 16 kHz. Also, in Fig.6(a) and 6(b), we see that the proposed method did not remove many estimation errors. These errors exist in the vicinity of direct current and higher than 16kHz. We see that both BSS-SVAD and Prop. can be accurately estimated IPD up to about 1.6kHz, which is used for direction perception. It can be seen that using only the high power frequency components contribute to the improvement of the estimation accuracy of ILD and IPD. These simulations showed that the Prop. is more effective for binaural sound separation in comparison to the BSS-SVAD.

In this paper, we propose a method to estimate ILD and IPD by using high power component in SV segment. The proposed method adjusts amplitude and phase of separated signal based on the estimated ILD and IPD. We also confirmed the effectiveness of the proposed method through objective evaluation experiments. Experimental results showed that the proposed method can accurately estimate ILD and IPD, although the estimation accuracy of the conventional BSS-SVAD.

## References

[1] K. Iida, M. Morimoto, Spatial acoustics, Corona Company, Tokyo, 2010.
[2] K. Nakadai, H. Okuno, H. Kitano, "Issues of auditory function in humanoids and sound localization by active audition," Journal of Artificial Intelligence Society, vol.18, no.2, pp.104-113, 2003.
[3] H. Hamada, "Binaural sound field reproduction system," Journal of Japan Acoustical Society, vol.48, no.4, pp.250-257, 1992.
[4] J. L. Flanagan, J. D. Johnston, R. Zahn, and G. W. Eiko, "Computer-streered microphone arrays for sound transduction in large morns," J. Acoust. Soc. Amer., vol.78, pp.1508-1518, 1985.

[5] N. Murata, Independent component analysis, Tokyo Denki University Press, Tokyo, 2004.
[6] H. Sawada, S. Araki and S. Makino, ”Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment,”IEEE Trans. Audio Speech Lang. Process.,vol.19, pp.516-527, 2010.
[7] A. Matsuda, A. Kawamura, Y. Iiguni, “Low computational two-channel blind source separation using single-voice activity segment for unknown number of sources,” Journal of the Japan acoustical society, vol.72, no.3, pp.115-122, 2016.
[8] R.O. Duda, et. al., Pattern Classification 2nd edition, Wiley Insterscience, 2000.
[9] MIT Media Lab, HRTF of KEMAR Dummy-Head, http://sound.media.mit.edu/resources/KEMAR.html.
[10] MedleyDB, A Dataset of Multitrack Audio for Music Research, http://medleydb.weebly.com.