第29　回信号処理シンポジウム
2014年11月11日～14日（京都）

チャネル別事前分布推定と両耳共通スペクトルゲインを用いた定位保持型バイノーラル音源分離

Localization-Preserved Binaural Source Separation Using Channel-Wise Target Prior Estimation and Equi-Binaural Spectral Gain

室田勇騎 [1], 北村大地 [2], 小山翔一 [3], 猿渡洋 [3], 中村哲 [1]
[1] 奈良先端科学技術大学院大学, [2] 総合研究大学院大学, [3] 東京大学

Yuki MUROTA [1], Daichi KITAMURA[2], Shoichi KOYAMA[3] ,
Hiroshi SARUWATARI[3] , Satoshi NAKAMURA[1]

[1] Nara Institute of Science and Technology, Nara, Japan
[2] The Graduate University for Advanced Studies (SOKENDAI), Kanagawa, Japan
[3] The University of Tokyo, Tokyo, Japan

Abstruct This paper addresses a new statistical model of binaural signals and its application to efficient binaural source separation. Binaural source separation is always required to retain a spatial cue of the separated sound, such as a head-related transfer function (HRTF). However, the direct use of an HRTF is not realistic because this information is normally not known in advance. To cope with this problem, first, we focus on the difference between signal probability density functions at both ears, which can be blindly estimated by using our previous work on higher-order statistics. Next, we derive a sound-localization-preserved generalized minimum mean-square error short-time spectral amplitude estimator. Objective and subjective experiments show the efficacy of the proposed method in terms of spatial quality.

1 Introduction

Audio signal separation has received much attention in signal-processing research, and many studies have been published in the last decade. Techniques for signal separation have been developed for many audio applications, including target speech enhancement for hearing-aid systems [1] and for controlling each source in a music tune in interactive 3D audio systems [2]–[5]. In this paper, we also address such an audio signal separation problem, especially focusing on a signal provided in a *binaural* format [6].

Compared with simple multichannel signal processing, binaural signal separation includes a relatively difficult task, namely, extraction of a specific sound while maintaining its spatial properties. This is because deterioration of the spatial quality of the separated sound has an adverse effect on human's 3D audio perception. Several methods have been proposed for binaural signal separation, mainly for blind speech separation and enhancement. To preserve a sound-localization cue such as the interaural level difference, these methods [7]–[10] apply an equi-binaural spectral gain to both the left and right ears of the listener, which can be calculated via, e.g., Wiener filtering (WF) [11] and the minimum mean-square error short-time spectral amplitude (MMSE-STSA) estimator [12], [13]. These methods have a drawback that they do not take account of binaural information. Ideally, the best way to enhance it is to explicitly use an important spatial cue, such as a head-related transfer function (HRTF) [6]. We have proposed an algorithm [14] to introduce a user's HRTF into a multichannel MMSE-STSA estimator but this method was not a blind system; the accurate measurement of the HRTF was required in advance, which is sometimes impossible in practice.

In another context of signal separation such as *informed source separation*, the authors have proposed a combination [15] of supervised nonnegative matrix factorization (SNMF) [16]–[19] and a prior-model-adapted generalized MMSE-STSA estimator [20] to deal with music signal separation. This method requires a certain supervision such as a music scale of the target instrument, but it efficiently extracts a target sound composed with an arbitrary melody from the observed monaural mixture. In this method, thanks to the higher-order statistics analysis, hidden parameters of the target statistical model can be estimated in each frequency subband.

Motivated by the above-mentioned prior work, in this paper, we propose a new spatial-cue-aware binaural signal separation algorithm without knowing the user's HRTF. The key idea and advantages of the proposed method are summarized as follows: (I) Instead of using an HRTF, we introduce a target statistical model to express the difference between user's left- and right-ear signals. For instance, if the target sound

is located on the left-hand side, the extracted signal in the left ear will obey a spiky probability density function (p.d.f.) but the right-ear signal will have a smooth p.d.f. because of many diffracted waves and the weak direct wave. (II) The statistical models for each ear can be accurately determined using only observable data [15]. Thus, the proposed strategy is HRTF-blind and a new attempt at establishing *a statistical HRTF approach*. (III) Using the same idea as in [10], an equi-binaural spectral gain is derived on the basis of statistical-HRTF-adapted generalized MMSE-STSA estimators. This avoids the marked deterioration of spatial quality.

2 Conventional Method

2.1 Single-Channel Music Signal Enhancement

In our previous work, we proposed single-channel music signal enhancement based on the generalized Bayesian estimator with automatic target prior adaptation [15]. Here, we use the SNMF-based dynamic interference spectrogram estimator and closed-form parameter estimation for the statistical model of the target signal based on higher-order statistics. The details are described below.

2.1.1 Music signal separation by generalized MMSE-STSA estimator with automatic target prior adaptation

We apply short-time Fourier analysis to the observed signal, which is a mixture of target and interference signals, to obtain the time-frequency-domain complex-valued signal

$$x(f,\tau)=s(f,\tau)+n(f,\tau), \tag{1}$$

where $x(f,\tau)$ is the observed signal, $s(f,\tau)$ is the target signal, $n(f,\tau)$ is the interference signal, f is the frequency bin number, and τ is the time-frame index.

For the generalized MMSE-STSA estimator, the amplitude spectrum of the target signal is estimated on the basis of the MMSE criterion under a certain target prior. The processed signal $\tilde{s}(f,\tau)$ via the generalized MMSE-STSA estimator is given by

$$\tilde{s}(f,\tau) = G(f,\tau)x(f,\tau), \tag{2}$$

$$G(f,\tau) = \frac{\sqrt{\nu(f,\tau)}}{\gamma(f,\tau)} \cdot \left(\frac{\Gamma(\rho+0.5)}{\Gamma(\rho)} \cdot \frac{\Phi(0.5-\rho,1,-\nu(f,\tau))}{\Phi(1-\rho,1,-\nu(f,\tau))} \right)^{1/\beta}, \tag{3}$$

where $\Gamma(\cdot)$ is the gamma function, $\Phi(a,b;k) = F_1(a,b;k)$ is the confluent hypergeometric function, β is the amplitude compression parameter, and

$$\nu(f,\tau)= \tilde{\gamma}(f,\tau)\tilde{\xi}(f,\tau)\left(1+\tilde{\xi}(f,\tau)\right)^{-1}. \tag{4}$$

Here, $\tilde{\xi}(f,\tau)$ and $\tilde{\gamma}(f,\tau)$ are the estimated a priori and a posteriori SNRs, respectively, which are defined as

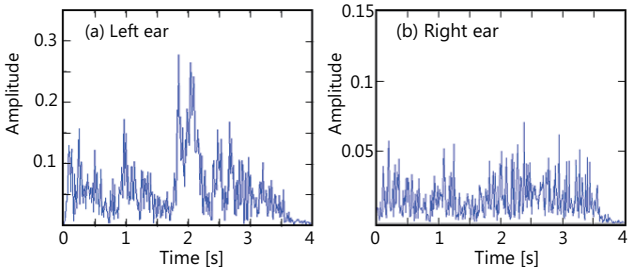

Fig. 1: Difference in signals at left and right ears.

$$\tilde{\xi}(f,\tau) = \alpha\tilde{\gamma}(f,\tau-1)G^2(f,\tau) + (1-\alpha)\max[\gamma(f,\tau)-1,0], \tag{5}$$

$$\tilde{\gamma}(f,\tau) = |x(f,\tau)|^2/P_{\tilde{n}}(f), \tag{6}$$

where $P_{\tilde{n}}(f)$ is the estimated interference power spectral density and α is the forgetting factor.

In the generalized MMSE-STSA estimator, the a priori statistical model of the target signal amplitude spectrum is set to the chi distribution

$$p(x) = 2\phi^{\rho}\Gamma(\rho)^{-1}x^{2\rho-1}\exp(-\phi x^2), \tag{7}$$

where $p(x)$ is the p.d.f. of signal x in the amplitude domain, $\phi = \rho/E\{|x|^2\}$, and ρ is the shape parameter. Here, $\rho = 1$ gives a Rayleigh distribution that corresponds to a Gaussian distribution in the time domain, and a smaller value of ρ corresponds to a super-Gaussian distribution signal.

In the generalized MMSE-STSA estimator, to calculate $\tilde{\gamma}(f,\tau)$, dynamic estimation is required if the interference signal is nonstationary, and estimation of the shape parameter ρ, which depends on the type of target signal, is also required.

2.1.2 Interference estimation by SNMF

The following equation represents the decomposition model of SNMF using the trained supervision components $F(f,k)$:

$$A(f,\tau) = |x(f,\tau)| \approx \sum_k F(f,k)V(k,\tau) + \sum_n H(f,n)U(n,\tau), \tag{8}$$

where $F(f,k)$ is a nonnegative element of the supervised basis matrix trained in advance, which comprises spectral patterns of the target signal as column vectors, $V(f,k)$ is a nonnegative element of an activation matrix that corresponds to $F(f,k)$, $H(f,n)$ represents a nonnegative element of the other basis matrix, which comprises residual spectral patterns that cannot be expressed by $\sum_k F(f,k)V(k,\tau)$, and $U(n,\tau)$ is a nonnegative element of the activation matrix that corresponds to $H(f,n)$. Moreover, k is the basis index of $F(f,k)$ and n is the basis index of $H(f,n)$. The supervised basis matrix can be trained using sample sounds of the target signal in the training process. Hence, ideally, $\sum_k F(f,k)V(k,\tau)$ represents the target signal components and $\sum_n H(f,n)U(n,\tau)$ represents the

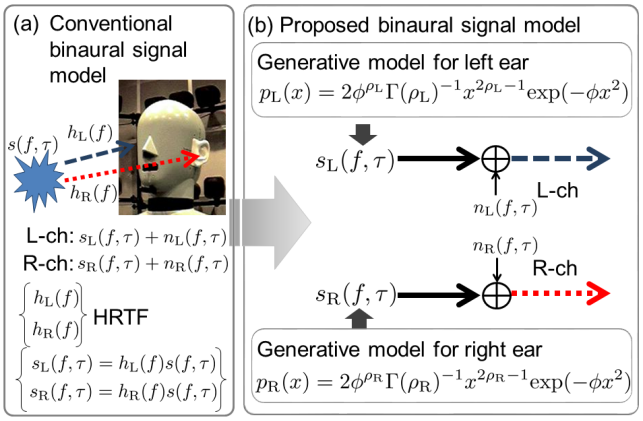

Fig. 2: Conventional and proposed binaural models.

other components different from the target signals after the decomposition. Thus, $(\sum_n H(f,n)U(n,\tau))^2$ is regarded as a good estimate of $P_{\tilde{N}}(f)$ in (6) in the time-frequency grids.

3 Proposed Method

3.1 Motivation and Strategy

In binaural source separation, it is desirable to use binaural cues such as an HRTF to improve the separation performance. However, this is difficult because we cannot obtain the HRTF of an unknown user blindly. Therefore, considering the HRTF from a statistical viewpoint, we efficiently express binaural cues using a statistical model based on the chi distribution. First, in the binaural signal, the p.d.fs. of the signals of each ear are assumed to be different by the influence of the difference in the signals arriving at each ear, and we can obtain a binaural cue based on the difference in these p.d.fs. For example, Fig. 1 shows amplitude of left- and right-ear signals of cello in 6 kHz subband, where the sound comes from the left-hand side. As shown in this figure, the left-ear signal contains several dominant components, which yields spiky p.d.f., but nothing in the right ear. Next, we introduce the chi distribution to represent the p.d.f. of each ear. By applying the target prior adaption algorithm [15] to each ear, it is possible to obtain suitable parameters of the p.d.fs. automatically. This means that we can convert the conventional deterministic HRTF estimation problem into a parameter estimation problem of the corresponding statistical model (see Fig. 2). This can also enable adaptation to unknown users.

However, there is a problem when we apply this strategy to a binaural source separation system. Generally speaking, a statistical-model-based source separation method (e.g., the generalized MMSE-STSA estimator) only provides the statistically fluctuating spectral gains for each of ears independently. However, the fluctuation of the gain function in interaural level differences at the left and right ears causes the deterioration in sound localization. To resolve this problem, we derive a new optimal spectral gain that minimizes the residual interference power in terms of the MMSE under the condition that the spectral gains are equivalent in both ears. Hereafter, we call this gain the *equi-binaural optimal spectral gain*.

3.2 Signal Mixture Model

We consider a mixing model with two inputs, i.e., two ears, and assume that the observed signal contains the target signal and an interference signal. Hereafter, the observed signal vector in the time-frequency domain, $\boldsymbol{x}(f,\tau) = [x_{\mathrm{L}}(f,\tau), x_{\mathrm{R}}(f,\tau)]^{\mathrm{T}}$, is given by

$$\boldsymbol{x}(f,\tau) = \boldsymbol{h}(f)s(f,\tau) + \boldsymbol{n}(f,\tau), \tag{9}$$

where $\boldsymbol{h}(f) = [h_{\mathrm{L}}(f), h_{\mathrm{R}}(f)]^{\mathrm{T}}$ is the column vector of the transfer functions between the target source and each ear, $s(f,\tau)$ is the target signal component, and $\boldsymbol{n}(f,\tau) = [n_{\mathrm{L}}(f,\tau), n_{\mathrm{R}}(f,\tau)]^{\mathrm{T}}$ is the column vector of the interference signal. Throughout this paper, the subscripts $*$ ($* = \{\mathrm{L},\mathrm{R}\}$) represent the signals obtained at the left and right ears, respectively.

3.3 Derivation of Equi-Binaural Optimal Spectral Gain

The derivation of the equi-binaural optimal spectral gain is described below. This is the extended version of [10] for a *generalized* cost function, and can be formulated as the minimization problem of the following error e:

$$\begin{aligned} e =& \mathrm{E}\Big[\{|h_{\mathrm{L}}(f)s(f,\tau)|^\beta - (G(f,\tau)|x_{\mathrm{L}}(f,\tau)|)^\beta\}^2 \\ &+ \{|h_{\mathrm{R}}(f)s(f,\tau)|^\beta - (G(f,\tau)|x_{\mathrm{R}}(f,\tau)|)^\beta\}^2\Big], \end{aligned} \tag{10}$$

where $G(f,\tau)$ is the equi-binaural spectral gain, which is considered as a variable. The optimization problem based on (10) is given by

$$\begin{aligned} G_{\mathrm{opt}}(f,\tau) = \underset{G(f,\tau)}{\mathrm{argmin}}\, \mathrm{E}\Big[&\{|h_{\mathrm{L}}(f)s(f,\tau)|^\beta - (G_{\mathrm{L}}(f,\tau)|x_{\mathrm{L}}(f,\tau)|)^\beta\}^2 \\ &+ \{|h_{\mathrm{R}}(f)s(f,\tau)|^\beta - (G_{\mathrm{R}}(f,\tau)|x_{\mathrm{R}}(f,\tau)|)^\beta\}^2 \\ &+ \{(G^\beta(f,\tau) - G_{\mathrm{L}}^\beta(f,\tau))|x_{\mathrm{L}}(f,\tau)|^\beta\}^2 \\ &+ \{(G^\beta(f,\tau) - G_{\mathrm{R}}^\beta(f,\tau))|x_{\mathrm{R}}(f,\tau)|^\beta\}^2 + 2C\Big], \end{aligned} \tag{11}$$

where $G_{\mathrm{opt}}(f,\tau)$ is the equi-binaural optimal spectral gain to be estimated, and $G_{\mathrm{L}}(f,\tau)$ and $G_{\mathrm{R}}(f,\tau)$ are individual spectral gains for the L and R ears, respectively, which are auxiliary parameters for calculating an approximate solution of $G_{\mathrm{opt}}(f,\tau)$ because the direct Bayesian estimation of $G_{\mathrm{opt}}(f,\tau)$ is difficult. In addition, C is related to the correlation between the estimation error and the observed signal in each channel when we estimate the target speech signals in the L and R ears using the parameters $G_{\mathrm{L}}(f,\tau)$ and $G_{\mathrm{R}}(f,\tau)$, and is defined by

$$\begin{aligned} C =& \{G^\beta(f,\tau) - G_{\mathrm{L}}^\beta(f,\tau)\} \\ &\cdot \{(G_{\mathrm{L}}(f,\tau)|x_{\mathrm{L}}(f,\tau)|)^\beta - |h_{\mathrm{L}}(f,\tau)s(f,\tau)|^\beta\}|x_{\mathrm{L}}(f,\tau)|^\beta \\ &+ \{G^\beta(f,\tau) - G_{\mathrm{R}}^\beta(f,\tau)\} \\ &\cdot \{(G_{\mathrm{R}}(f,\tau)|x_{\mathrm{R}}(f,\tau)|)^\beta - |h_{\mathrm{R}}(f,\tau)s(f,\tau)|^\beta\}|x_{\mathrm{R}}(f,\tau)|^\beta. \end{aligned} \tag{12}$$

We discuss the minimization of (11). First, the 1st and 2nd terms on the right-hand side correspond to the problem of target signal estimation in each ear. These terms can be mini-

mized if we obtain the optimal values of $G_{\mathrm{L}}(f,\tau)$ and $G_{\mathrm{R}}(f,\tau)$ using the generalized MMSE-STSA estimator described in Sect. 2. Next, C in the 5th term on the right-hand side can be disregarded if the parameters $G_{\mathrm{L}}(f,\tau)$ and $G_{\mathrm{R}}(f,\tau)$ provide an accurate estimate of the target signals by approximately considering C to be negligible. Hence, the residual 3rd and 4th terms, i.e., $\{(G^{\beta}(f,\tau)-G^{\beta}_{\mathrm{L}}(f,\tau))|x_{\mathrm{L}}(f,\tau)|^{\beta}\}^2+\{(G^{\beta}(f,\tau)-G^{\beta}_{\mathrm{R}}(f,\tau))|x_{\mathrm{R}}(f,\tau)|^{\beta}\}^2$, should be minimized. This problem can be formulated as

$$\begin{aligned}&G_{\mathrm{opt}}(f,\tau)\\&\simeq \underset{G(f,\tau)}{\operatorname{argmin}}\,\mathrm{E}\Big[\{(G^{\beta}(f,\tau)-G^{\beta}_{\mathrm{L_{opt}}}(f,\tau))|x_{\mathrm{L}}(f,\tau)|^{\beta}\}^2\\&\quad+\{(G^{\beta}(f,\tau)-G^{\beta}_{\mathrm{R_{opt}}}(f,\tau))|x_{\mathrm{R}}(f,\tau)|^{\beta}\}^2\Big],\end{aligned}\tag{13}$$

subject to

$$G_{\mathrm{L_{opt}}}(f,\tau)=\underset{G_{\mathrm{L}}(f,\tau)}{\operatorname{argmin}}\,\mathrm{E}\Big[\{|h_{\mathrm{L}}(f)s(f,\tau)|^{\beta}-(G_{\mathrm{L}}(f,\tau)|x_{\mathrm{L}}(f,\tau)|)^{\beta}\}^2\Big],\tag{14}$$

$$G_{\mathrm{R_{opt}}}(f,\tau)=\underset{G_{\mathrm{R}}(f,\tau)}{\operatorname{argmin}}\,\mathrm{E}\Big[\{|h_{\mathrm{R}}(f)s(f,\tau)|^{\beta}-(G_{\mathrm{R}}(f,\tau)|x_{\mathrm{R}}(f,\tau)|)^{\beta}\}^2\Big],\tag{15}$$

where $G_{\mathrm{Lopt}}(f,\tau)$ and $G_{\mathrm{Ropt}}(f,\tau)$ are the L- and R-ear optimal spectral gains, respectively.

To solve (13), we first obtain $G_{\mathrm{Lopt}}(f,\tau)$ and $G_{\mathrm{Ropt}}(f,\tau)$ from the generalized MMSE-STSA estimator in (14) and (15), then by substituting them into (13), we solve the following equation in $G(f,\tau)$:

$$\begin{aligned}\frac{\partial e}{\partial G(f,\tau)}&=G^{\beta}(f,\tau)|x_{\mathrm{L}}(f,\tau)|^{2\beta}-G^{\beta}_{\mathrm{Lopt}}(f,\tau)|x_{\mathrm{L}}(f,\tau)|^{2\beta}\\&\quad+G^{\beta}(f,\tau)|x_{\mathrm{R}}(f,\tau)|^{2\beta}-G^{\beta}_{\mathrm{Ropt}}(f,\tau)|x_{\mathrm{R}}(f,\tau)|^{2\beta}\\&=0.\end{aligned}\tag{16}$$

The solution of (16) is given by

$$\begin{aligned}&G_{\mathrm{opt}}(f,\tau)\\&=\left(\frac{G^{\beta}_{\mathrm{Lopt}}(f,\tau)|x_{\mathrm{L}}(f,\tau)|^{2\beta}+G^{\beta}_{\mathrm{Ropt}}(f,\tau)|x_{\mathrm{R}}(f,\tau)|^{2\beta}}{|x_{\mathrm{L}}(f,\tau)|^{2\beta}+|x_{\mathrm{R}}(f,\tau)|^{2\beta}}\right)^{1/\beta}.\end{aligned}\tag{17}$$

3.4 Shape Parameter and Kurtosis

In (17), we need to calculate $G_{\mathrm{Lopt}}(f,\tau)$ and $G_{\mathrm{Ropt}}(f,\tau)$, which include a shape parameter ρ that should represent the a priori distribution of the target signal. In Sects. 3.4 and 3.5, we describe how to blindly estimate ρ.

Regarding the chi distribution $p(x)$ in (7), the mth-order moment can be written as

$$\mu_m(x)=\int_0^{\infty}x^m p(x)dx=\frac{\Gamma(\rho+\frac{m}{2})}{\Gamma(\rho)}\phi^{-\frac{m}{2}}.\tag{18}$$

Then, the kurtosis of the chi distribution is calculated as

$$\mathrm{kurt}=\mu_4(x)/\mu_2^2(x)=(\rho+1)/\rho.\tag{19}$$

Therefore, the shape parameter ρ is given by

$$\rho=(\mathrm{kurt}-1)^{-1}.\tag{20}$$

From this relation, the shape parameter of the target signal can be estimated by obtaining its amplitude-domain kurtosis value. In general, however, it is difficult to directly estimate the kurtosis of a target signal because of its contamination by additive interference signals.

3.5 Estimation of Hidden Target Kurtosis and Gain Function

In our previous work, we proposed an algorithm for target kurtosis estimation in additive signals, which can be derived from the closed-form relation in higher-order statistics. In this algorithm, the resultant kurtosis of the target amplitude spectrum is estimated as

$$\begin{aligned}\mathrm{kurt}_*=&\left(\mu_4(A_*)-\mu_4\left(\sum_n(HU)_*\right)\right.\\&\left.+4\mu_2^2(\sum_n(HU)_*)-4\mu_2(A_*)\mu_2(\sum_n(HU)_*)\right)\\&\cdot\left(\mu_2^2(A_*)+\mu_2^2(\sum_n(HU)_*)-2\mu_2(A_*)\mu_2(\sum_n(HU)_*)\right)^{-1},\end{aligned}\tag{21}$$

where we ignore the indexes f and τ for saving the space. For the detailed derivation of (21), see Ref. [15].

The shape parameter of the target signal p.d.f. at each ear can be estimated using the kurtosis and (20). Therefore, the equi-binaural optimal spectral gain estimated by the proposed method is obtained as follows by substituting (2) into (17):

$$\begin{aligned}&\tilde{G}_{\mathrm{opt}}(f,\tau)\\&=\left\{\frac{|x_{\mathrm{L}}(f,\tau)|^{2\beta}(\tilde{\nu}_{\mathrm{L}}(f,\tau))^{\beta/2}\Gamma\left((\mathrm{kurt}_{\mathrm{L}}-1)^{-1}+0.5\right)}{\{|x_{\mathrm{L}}(f,\tau)|^{2^{\beta}}+|x_{\mathrm{R}}(f,\tau)|^{2\beta}\}\hat{\gamma}^{\beta}_{\mathrm{L}}(f,\tau)\Gamma\left((\mathrm{kurt}_{\mathrm{L}}-1)^{-1}\right)}\right.\\&\quad\cdot\frac{\Phi(0.5-(\mathrm{kurt}_{\mathrm{L}}-1)^{-1},1,-\tilde{\nu}_{\mathrm{L}}(f,\tau))}{\Phi(1-(\mathrm{kurt}_{\mathrm{L}}-1)^{-1},1,-\tilde{\nu}_{\mathrm{L}}(f,\tau))}\\&+\frac{|x_{\mathrm{R}}(f,\tau)|^{2\beta}(\tilde{\nu}_{\mathrm{R}}(f,\tau))^{\beta/2}\Gamma\left((\mathrm{kurt}_{\mathrm{R}}-1)^{-1}+0.5\right)}{\{|x_{\mathrm{L}}(f,\tau)|^{2\beta}+|x_{\mathrm{R}}(f,\tau)|^{2\beta}\}\hat{\gamma}^{\beta}_{\mathrm{R}}(f,\tau)\Gamma\left((\mathrm{kurt}_{\mathrm{R}}-1)^{-1}\right)}\\&\quad\left.\cdot\frac{\Phi(0.5-(\mathrm{kurt}_{\mathrm{R}}-1)^{-1},1,-\tilde{\nu}_{\mathrm{R}}(f,\tau))}{\Phi(1-(\mathrm{kurt}_{\mathrm{R}}-1)^{-1},1,-\tilde{\nu}_{\mathrm{R}}(f,\tau))}\right\}^{1/\beta}.\end{aligned}\tag{22}$$

The final output is given by $\tilde{s}_*(f,\tau)=\tilde{G}_{\mathrm{opt}}(f,\tau)x_*(f,\tau)$.

4 Evaluation Experiments

4.1 Experimental Conditions

In this experiment, we used four binaural instrumental signals, namely, an oboe, clarinet, cello, and piano, where the target instrument $s(f,\tau)$ is the oboe (each melody part is depicted in [15]). These signals were artificially generated by a MIDI synthesizer and the directions of arrival of these signals were set from −90° to 90° with 15° intervals by using the

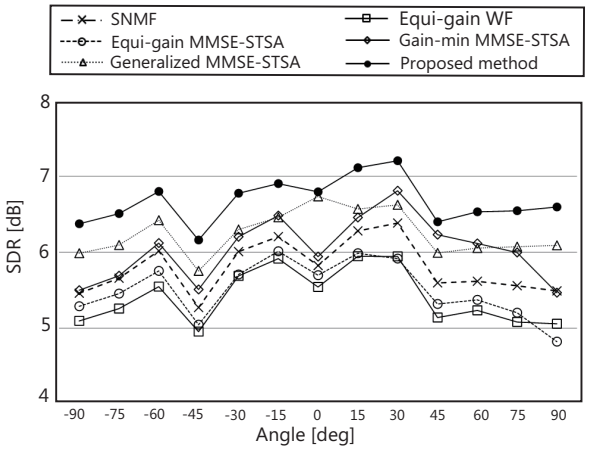

Fig. 3: Average SDRs for each method and each direction.

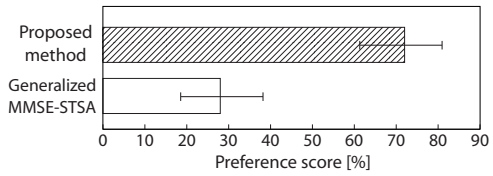

Fig. 4: Result of XAB test for evaluation of spatial quality.

corresponding HRTF $\boldsymbol{h}(f)$. The HRTFs were selected from the open database of the "Samurai" dummy head [21]. The observed signals were produced by mixing two sources selected from the target signal and the other three signals with the same power. In this observed signal, the target and interference signals were located in the same direction. In the estimation of the interference signal using SNMF, we used artificial clean MIDI sounds of the target instrument without an HRTF as supervision for the training process. The training sounds contained two octave notes that covered all the notes of the target signal in the observed signal. The sampling frequency of all signals was 44.1 kHz. Spectrograms were computed using a 92-ms-long rectangular window with an 11-ms-long overlap shift. Moreover, the number of trained bases was 100 and the number of other bases was 50. The forgetting factor α was 0.97, and the amplitude compression parameter β was 1.0.

4.2 Objective Experiment

In the objective experiment, we compared six methods, i.e., SNMF (**SNMF**) [19], the equi-binaural spectral gain via WF (**Equi-gain WF**) [9], the equi-binaural spectral gain via the MMSE-STSA estimator (**Equi-gain MMSE-STSA**) [10], the minimum-gain-based MMSE-STSA estimator (**Gain-min MMSE-STSA**) [8], the generalized MMSE-STSA estimators independently applied to each ear (**Generalized MMSE-STSA**), and the equi-binaural spectral gain via the generalized MMSE-STSA estimator (**Proposed method**). Every method uses the same SNMF as the interference estimator. We used the signal-to-distortion ratio (SDR) defined in [22] as the evaluation score. The SDR indicates the overall quality of the separated target sound, and is high in the case of high separation, low artificial distortion, and low spatial distortion.

Figure 3 shows the average SDRs for each method and each direction. From these results, we can confirm that the separation performance of the proposed method is better than those of the other methods. This result indicates the efficacy of introducing the flexible a priori statistical model of the target signal and equi-binaural spectral gain. The simple MMSE-STSA estimators (Equi-gain MMSE-STSA and Gain-min MMSE-STSA) also assume the fixed a priori model of the Gaussian distribution but the assumption is not appropriate for representing a music signal and a spatial difference between both ears. In contrast, the proposed method automatically chooses a spikier p.d.f. ($\rho \ll 1$) for the ear closer to the source location and a smoother p.d.f. ($\rho \simeq 1$) for the opposite ear. These p.d.fs. match the binaural target.

4.3 Subjective Experiment

We next conducted a subjective test to evaluate the performance of the proposed method, focusing on the human impression of the separated signal from the viewpoint of spatial quality. In the subjective experiment, we employed the XAB method and compared two methods, i.e., Generalized MMSE-STSA and Proposed method. The participants in the experiment comprised four males and two females.

Figure 4 shows the result of the subjective experiment, which indicates that the proposed method using the equi-binaural spectral gain markedly outperforms Generalized MMSE-STSA. Therefore, we confirmed the effectiveness of using the equi-binaural spectral gain to improve the spatial quality.

5 Conclusion

In this paper, to address the effect of statistical models for both ears on binaural signal source separation, we applied the generalized MMSE-STSA estimator with automatic prior adaptation to a binaural signal. The proposed method of binaural signal separation using equi-binaural spectral gain can also improve the sound-localization properties. From the results of and subjective experiments, it was found that the proposed method outperforms conventional methods from the viewpoint of separation performance and sound-localization preservation.

References

[1] J. Benesty, S. Makino, J. Chen, *"Speech Enhancement,"* Springer Press, Springer Science+Business Media, Berlin, 2005.

[2] N. Kamado, H. Nawata, H. Saruwatari, K. Shikano, "Interactive controller for audio object localization based on spatial representative vector operation," *Proc. 2010 International Workshop on Acoustic Echo and Noise Control (IWAENC2010)*, 2010.

[3] J. Herre, H. Purnhagen, J. koppens, O. Hellmuth, J. Engdegård, J. Hilpert, L. Villemoes, L. Terentiv,

C. Falch, A. Hölzer, M. L. Valero, B. Resch, H. Mundt, H.-O. Oh, "MPEG spatial audio object coding-the ISO/MPEG standard for efficient coding of interactive audio scenes," *J. Audio Eng. Soc.*, vol.60, no.9, pp.655–673, 2012.

[4] N. Kamado, M. Hirata, H. Saruwatari, K. Shikano, "Object-based stereo up-mixer for wave field synthesis based on spatial information clustering," *Proc. 20th European Signal Processing Conference (EUSIPCO2012)*, pp.594–598, 2012.

[5] S. Koyama, K. Furuya, Y. Hiwasaki, and Y. Haneda, "Design of transform filter for reproducing arbitrary shifted sound field using phase-shift of spatio-temporal frequency," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2012)*, pp.381–384, 2012.

[6] J. Blauert, *"Spatial Hearing,"* The MIT Press, Cambrige Massachusetts London, England, 1996.

[7] T. van den Bogaert, S. Doclo, J. Wouters, M. Moonen, "The effect of multimicrophone noise reduction systems on sound source localization by users of binaural hearing aids," *Journal of the Acoustical Society of America*, vol.124, no.1, pp.484–497, 2008.

[8] A. H. Kamkar-Parsi, M. Bouchard, "Improved noise power spectrum density estimation for binaural hearing aids operating in a diffuse noise field environment," *IEEE Trans. Audio, Speech and Lang. Process.*, vol.17, no.4, pp.521–533, 2009.

[9] K. Reindl, Y. Zheng, W. Kellermann, "Speech enhancement for binaural hearing aids based on blind source separation," *Proc. 2010 International Symposium on Communications, Control and Signal Processing (ISCCSP2010)*, 2010.

[10] H. Saruwatari, M. Go, R. Okamoto, K. Shikano, H. Hosoi, "Binaural hearing aid using sound-localization-preserved MMSE STSA estimator with ICA-based noise estimation," *Proc. 2010 International Workshop on Acoustic Echo and Noise Control (IWAENC2010)*, 2010.

[11] P. C. Loizou, *"Speech Enhancement Theory and Practice,"* CRC Press, Taylor & Francis Group, FL, 2007.

[12] Y. Ephraim, D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol.32, no.6, pp.1109–1121, 1984.

[13] Y. Ephraim, D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol.32, no.2, pp.443–445, 1985.

[14] F. Mustiere, M. Bouchard, H. Najaf-Zadeh, R. Pichevar, L. Thibault, H. Saruwatari, "Design of multichannel frequency domain statistical-based enhancement systems preserving spatial cues via spectral distances minimization," *Signal Processing (Elsevier)*, vol.93, no.1, pp.321–325, 2013.

[15] Y. Murota, D. Kitamura, S. Nakai, H. Saruwatari, S. Nakamura, K. Shikano, Y. Takahashi, K. Kondo, "Music signal separation based on bayesian spectral amplitude estimator with automatic target prior adaptation," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2014)*, pp.7490–7494, 2014.

[16] P. Smaragdis, B. Raj, M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," *Proc. 7th International Conference on Independent Component Analysis and Signal Separation*, pp.414–421, 2007.

[17] S.-Y. Jeong, K. Kim, J.-H. Jeong, K.-C. Oh, "Semi-blind disjoint non-negative matrix factorization for extracting target source from single channel noisy mixture," *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA2009)*, pp.73–76, 2009.

[18] E. M. Grais, H. Erdogan, "Discriminative nonnegative dictionary learning using cross-coherence penalties for single channel source separation," *Proc. INTERSPEECH 2013*, pp.808–812, 2013.

[19] D. Kitamura, H. Saruwatari, K. Yagi, K. Shikano, Y. Takahashi, K. Kondo, "Music signal separation based on supervised nonnegative matrix factorization with orthogonality and maximum-divergence penalties," *IEICE Trans. Fundamentals of Electronics, Communications and Computer Sciences*, vol.E97-A, no.5, pp.1113–1118, 2014.

[20] C. Breithaupt, M. Krawczyk, R. Martin, "Parameterized MMSE spectral magnitude estimation for the enhancement of noisy speech," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.4037–4040, 2008.

[21] S. Shimada, K. Sugiyama, H. Hokari *"Head-Related Transfer Function,"* Cyber Publishing Center Press, Cyber Creative Institute, Japan, 2014 (in Japanese).

[22] E. Vincent, R. Gribonval, C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech and Language Processing*, vol.14, no.4, pp.1462–1469, 2006.