BLIND SEPARATION OF BINAURAL SOUND MIXTURES USING SIMO-MODEL-BASED INDEPENDENT COMPONENT ANALYSIS

Tomoya TAKATANI, Tsuyoki NISHIKAWA, Hiroshi SARUWATARI, and Kiyohiro SHIKANO

Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma-shi, Nara, 630-0192, JAPAN
E-mail: {tomoya-t, tsuyo-ni, sawatari, shikano}@is.aist-nara.ac.jp

ABSTRACT

High-fidelity blind audio signal separation is addressed adopting the extended ICA algorithm, single-input multiple-output (SIMO)-model-based ICA. The SIMO-ICA consists of multiple ICA parts and a fidelity controller, and each ICA runs in parallel under fidelity control of the entire separation system. SIMO-ICA can separate the mixed signals, not into monaural source signals, but into SIMO-model-based signals from independent sources as they are at the microphones. Thus, the separated signals of the SIMO-ICA can maintain the spatial qualities of each sound source. In this paper, we apply the SIMO-ICA to blind separation problem of mixed binaural sounds including the effect of the head-related transfer function (HRTF). The experimental results reveal that the performance of the proposed SIMO-ICA is superior to that of the conventional ICA-based method, and the separated signals of SIMO-ICA maintain the spatial qualities of each sound source.

1. INTRODUCTION

Blind source separation (BSS) is the approach taken to estimate original source signals using only the information of the mixed signals observed in each input channel. This technique is applicable to high-quality hands-free telecommunication systems. In recent works of BSS based on independent component analysis (ICA) [1], various methods have been proposed to deal with a means of separation of acoustic sounds which corresponds to the convolutive mixture case [2]–[4]. However, the conventional ICA-based BSS approaches are basically means of extracting each of the independent sound sources as a *monaural* signal, and consequently they have a serious drawback in that the separated sounds cannot maintain information about the directivity, localization, or spatial qualities of each sound source. This prevents any BSS method from being applied to binaural signal processing [5] or high-fidelity sound reproduction systems [6].

In order to solve the above-mentioned fundamental problems, we have proposed the high-fidelity BSS using Single-Input Multiple-Output (SIMO)-model-based ICA [7], in which the convolutive mixtures of acoustic signals are decomposed into the SIMO components. Here the term "SIMO" represents the specific transmission system in which the input is a single source signal and the outputs are its transmitted signals observed at multiple microphones. The SIMO-ICA consists of multiple ICA parts and a fidelity controller, and each ICA runs in parallel under the fidelity control of the entire separation system. The SIMO-ICA can separate the mixed signals, not into monaural source signals but into SIMO-model-based signals from independent sources as they are at the microphones. Thus, the separated signals of the SIMO-ICA can maintain the spatial qualities of each sound source.

In this paper, we apply the SIMO-ICA to the blind separation problem of binaural sounds including the effect of the head-related transfer function (HRTF) [5]. We carried out the separation experiments of binaural sounds recorded using a head and torso simulator (HATS) under a reverberant condition. From the results, it is revealed that the performance of separated signals of the SIMO-ICA is superior to that of the conventional ICA method. Also, from the shape of HRTF estimated by the SIMO-ICA, it can be shown that the output signals of SIMO-ICA maintain information about spatial qualities of each sound source.

2. MIXING PROCESS AND CONVENTIONAL BSS

2.1. Mixing Process

In this study, the number of microphones is K and the number of multiple sound sources is L. In general, the observed signals in which multiple source signals are mixed linearly are expressed as

$$\boldsymbol{x}(t) = \sum_{n=0}^{N-1} \boldsymbol{a}(n)\boldsymbol{s}(t-n) = \boldsymbol{A}(z)\boldsymbol{s}(t), \tag{1}$$

where $\boldsymbol{s}(t) = [s_1(t), \cdots, s_L(t)]^{\mathrm{T}}$ is the source signal vector and $\boldsymbol{x}(t) = [x_1(t), \cdots, x_K(t)]^{\mathrm{T}}$ is the observed signal vector. Also, $\boldsymbol{a}(n) = [a_{kl}(n)]_{kl}$ is the mixing filter matrix with the length of N, and $\boldsymbol{A}(z) = [A_{kl}(z)]_{kl} = [\sum_{n=0}^{N-1} a_{kl}(n)z^{-n}]_{kl}$ is the z-transform of $\boldsymbol{a}(n)$, where z^{-1} is used as the unit-delay operator, i.e., $z^{-n} \cdot x(t) = x(t-n)$, a_{kl} is the impulse response between the k-th microphone and the l-th sound source, and $[X]_{ij}$ denotes the matrix which includes the element X in the i-th row and the j-th column. Hereafter, we only deal with the case of $K = L$ in this paper.

2.2. Conventional ICA-based BSS Method

In the BSS method, we consider the time-domain ICA (TDICA), in which each element of the separation matrix is represented as an FIR filter. In the TDICA, we optimize the separation matrix by using only the fullband observed signals without subband processing. The separated signal $\boldsymbol{y}(t) = [y_1(t), \cdots, y_L(t)]^{\mathrm{T}}$ is expressed as

$$\boldsymbol{y}(t) = \sum_{n=0}^{D-1} \boldsymbol{w}(n)\boldsymbol{x}(t-n), \tag{2}$$

where $\boldsymbol{w}(n)$ is the separation filter matrix and D is the filter length of $\boldsymbol{w}(n)$. In our study, the separation filter matrix is optimized by

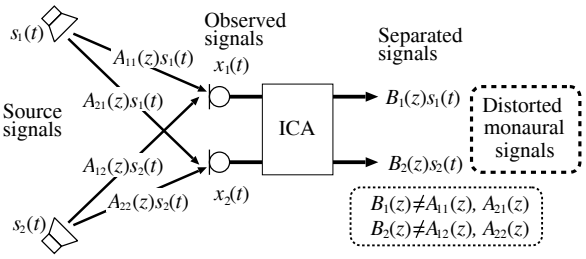

Fig. 1. Input and output relations in conventional ICA.

minimizing the Kullback-Leibler divergence (KLD) between the joint probability density function (PDF) of $\boldsymbol{y}(t)$ and the product of marginal PDFs of $y_l(t)$. The iterative learning rule is given by [8]

$$\begin{aligned}\boldsymbol{w}^{[j+1]}(n) &= \boldsymbol{w}^{[j]}(n) \\ &\quad - \alpha \sum_{d=0}^{D-1} \left\{ \text{off-diag} \left\langle \boldsymbol{\varphi}(\boldsymbol{y}^{[j]}(t)) \boldsymbol{y}^{[j]}(t-n+d)^{\mathrm{T}} \right\rangle_t \right\} \\ &\quad \cdot \boldsymbol{w}^{[j]}(d), \end{aligned} \tag{3}$$

where α is the step-size parameter, the superscript $[j]$ is used to express the value of the j-th step in the iterations, $\langle\cdot\rangle_t$ denotes the time-averaging operator, and off-diag$\boldsymbol{X}$ is the operation for setting every diagonal element of the matrix $\boldsymbol{X}$ to be zero. Also, $\boldsymbol{\varphi}(\cdot)$ is the nonlinear vector function, e.g., the l-th element is set to be $\tanh(y_l(t))$.

2.3. Problems in Conventional ICA

The conventional ICA is basically a means of extracting each of the independent sound sources as a monaural signal. In addition, the quality of the separated sound cannot be guaranteed, i.e., the separated signals can possibly include spectral distortions because the modified separated signals which convolved with arbitrary linear filters are still mutually independent (see Fig. 1). Therefore, the conventional ICA has a serious drawback in that the separated sounds cannot maintain information about the directivity, localization, or spatial qualities of each sound source. In order to resolve the problems, particularly for the sound quality, Matsuoka et al. have proposed a modified ICA based on the Minimal Distortion Principle [9]. However, this method is valid for only monaural outputs, and the fidelity of the output signals as SIMO-model-based signals cannot be guaranteed.

3. PROPOSED ALGORITHM: SIMO-ICA

In order to solve the above-mentioned fundamental problems, we have proposed a new blind separation framework that SIMO-model-based acoustic signals are separated by the extended ICA algorithm, SIMO-ICA [7]. The SIMO-ICA consists of $(l-1)$ ICA parts and a *fidelity controller*, and each ICA runs in parallel under the fidelity control of the entire separation system (see Fig. 2). The separated signals of the l-th ICA ($l = 1, \cdots, L-1$) in the SIMO-ICA are defined by

$$\boldsymbol{y}_{(\mathrm{ICA}l)}(t) = [y_k^{(\mathrm{ICA}l)}(t)]_{k1} = \sum_{n=0}^{D-1} \boldsymbol{w}_{(\mathrm{ICA}l)}(n)\boldsymbol{x}(t-n), \tag{4}$$

where $\boldsymbol{w}_{(\mathrm{ICA}l)}(n)$ is the separation filter matrix in the l-th ICA. Regarding the fidelity controller, we calculate the following signal vector, in which all of the elements are to be mutually independent,

$$\begin{aligned}\boldsymbol{y}_{(\mathrm{ICA}L)}(t) &= \boldsymbol{x}(t-\frac{D}{2}) - \sum_{l=1}^{L-1} \boldsymbol{y}_{(\mathrm{ICA}l)}(t) \\ &= \sum_{n=0}^{D-1} \boldsymbol{w}_{(\mathrm{ICA}L)}(n)\boldsymbol{x}(t-n). \end{aligned} \tag{5}$$

To explicitly show the meaning of the fidelity controller, we rewrite Eq. (5) as

$$\sum_{l=1}^{L} \boldsymbol{y}_{(\mathrm{ICA}l)}(t) - \boldsymbol{x}(t-D/2) = [0]_{k1}. \tag{6}$$

Equation (6) means a constraint to force the sum of all of the ICAs' output vectors $\sum_{l=1}^{L} \boldsymbol{y}_{(\mathrm{ICA}l)}(t)$ to be the sum of all of the SIMO components $[\sum_{l=1}^{L} A_{kl}(z)s_l(t-D/2)]_{k1} (= \boldsymbol{x}(t-D/2))$. Here the delay of $D/2$ is used as to deal with nonminimum phase systems. If the independent sound sources are separated by Eq. (4), and simultaneously the signals obtained by Eq. (5) are also mutually independent, then the output signals converge on unique solutions, up to the permutation, as

$$\boldsymbol{y}_{(\mathrm{ICA}l)}(t) = \mathrm{diag}\left[\boldsymbol{A}(z)\boldsymbol{P}_l^{\mathrm{T}}\right] \boldsymbol{P}_l \boldsymbol{s}(t-D/2), \tag{7}$$

where $\boldsymbol{P}_l$ ($l = 1, \ldots, L$) are exclusively-selected permutation matrices which satisfy

$$\sum_{l=1}^{L} \boldsymbol{P}_l = [1]_{ij}. \tag{8}$$

Obviously the solutions given by Eq. (7) provide necessary and sufficient SIMO components, $A_{kl}(z)s_l(t-D/2)$, for each l-th source. However, the condition Eq. (8) allows multiple possibilities for the combination of $\boldsymbol{P}_l$. For example, one possibility is shown in Fig. 2 and this corresponds to

$$\boldsymbol{P}_l = [\delta_{im(k,l)}]_{ki}, \tag{9}$$

where δ_{ij} is Kronecker's delta function, and

$$m(k,l) = \begin{cases} k+l-1 & (k+l-1 \leq L) \\ k+l-1-L & (k+l-1 > L) \end{cases}. \tag{10}$$

In this case, Eq. (7) yields

$$\boldsymbol{y}_{(\mathrm{ICA}l)}(t) = [A_{km(k,l)}s_{m(k,l)}(t-D/2)]_{k1} \quad (l = 1, \cdots, L). \tag{11}$$

In order to obtain Eq. (7), the natural gradient [4] of KLD of Eq. (5) with respect to $\boldsymbol{w}_{(\mathrm{ICA}l)}(n)$ should be added to the iterative learning rule of the separation filter in the l-th ICA ($l = 1, \cdots, L-1$). The new iterative algorithm of the l-th ICA part ($l = 1, \cdots, L-1$) in SIMO-ICA is given as

$$\begin{aligned}&\boldsymbol{w}_{(\mathrm{ICA}l)}^{[j+1]}(n) \\ &= \boldsymbol{w}_{(\mathrm{ICA}l)}^{[j]}(n) - \alpha \sum_{d=0}^{D-1} \Bigg[\Bigg\{ \text{off-diag} \left\langle \boldsymbol{\varphi}\big(\boldsymbol{y}_{(\mathrm{ICA}l)}^{[j]}(t)\big) \right. \\ &\quad \left. \boldsymbol{y}_{(\mathrm{ICA}l)}^{[j]}(t-n+d)^{\mathrm{T}} \right\rangle_t \Bigg\} \boldsymbol{w}_{(\mathrm{ICA}l)}^{[j]}(d) \\ &\quad - \Bigg\{ \text{off-diag} \left\langle \boldsymbol{\varphi}\big(\boldsymbol{x}(t-\frac{D}{2}) - \sum_{l=1}^{L-1} \boldsymbol{y}_{(\mathrm{ICA}l)}^{[j]}(t)\big) \right. \\ &\quad \left. \Big(\boldsymbol{x}(t-n+d-\frac{D}{2}) - \sum_{l=1}^{L-1} \boldsymbol{y}_{(\mathrm{ICA}l)}^{[j]}(t-n+d)^{\mathrm{T}}\Big) \right\rangle_t \Bigg\} \\ &\quad \Big(\boldsymbol{I}\delta(d-\frac{D}{2}) - \sum_{l=1}^{L-1} \boldsymbol{w}_{(\mathrm{ICA}l)}^{[j]}(d)\Big) \Bigg], \end{aligned} \tag{12}$$

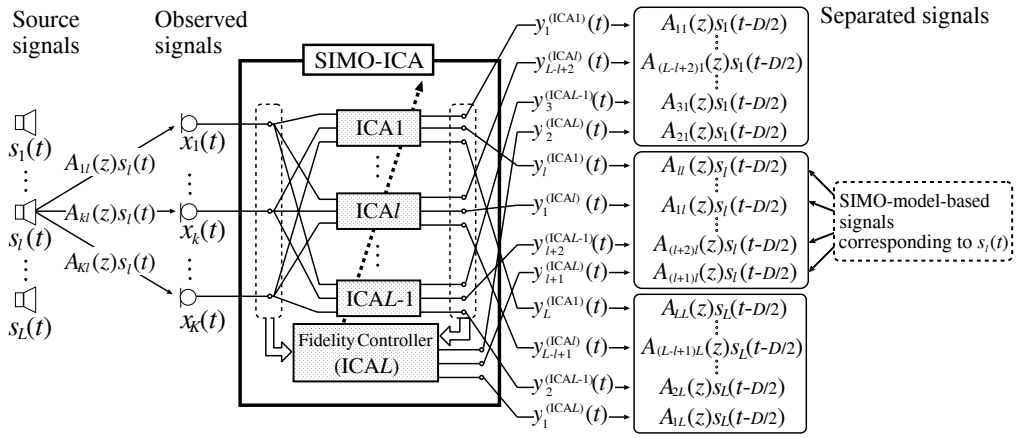

Fig. 2. Example of input and output relations in the proposed SIMO-ICA, where permutation matrices $\boldsymbol{P}_l$ is given by Eq. (9).

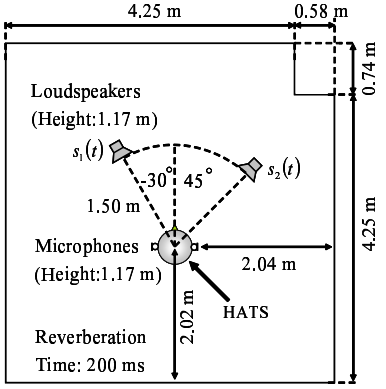

Fig. 3. Layout of reverberant room used in experiments.

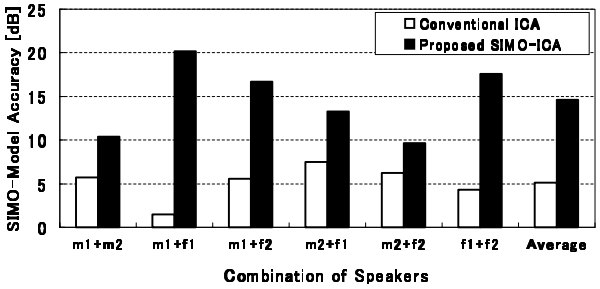

Fig. 4. Results of SIMO-model accuracy in separation experiments of binaural sounds recorded using HATS. The labels "m1" and "m2" mean two male speakers, and "f1" and "f2" mean two female speakers.

where $\delta(n)$ is a delta function, where $\delta(0) = 1$ and $\delta(n) = 0$ $(n \neq 0)$. In Eq. (12), the updating $\boldsymbol{w}_{(\mathrm{ICA}l)}(n)$ for all l should be simultaneously performed in parallel because each iterative equation is associated with the others via $\sum_{l=1}^{L-1} \boldsymbol{y}^{[j]}_{(\mathrm{ICA}l)}(t)$. Also, the initial values of $\boldsymbol{w}_{(\mathrm{ICA}l)}(n)$ for all l should be different.

4. EXPERIMENTS AND RESULTS

4.1. Conditions for Experiments

We carried out binaural sound separation experiments using speech signals convolved with impulse responses recorded using HATS (Brüel & Kjær) under the experimental room as shown in Fig. 3. The speech signals are assumed to arrive from two directions, $-30°$ and $45°$. The distance between HATS and the loudspeakers is 1.5 m. Two kinds of sentences, spoken by two male and two female speakers, are used as the original speech samples. Using these sentences, we obtain 6 combinations. The sampling frequency is 8 kHz and the length of speech is limited to 3 seconds. The length of $\boldsymbol{w}(n)$ is 512, and the initial values are inverse filters of HRTFs whose directions of sources are $\pm 60°$. The number of iterations in ICA is 5000. Regarding the conventional ICA for comparison, we used the nonholonomic ICA [8]. The step-size parameter α is changed from 1.0×10^{-8} to 2.0×10^{-6} to find the optima. SIMO-model accuracy (SA) is used as an evaluation score. The SA is defined as

$$\mathrm{SA} = \frac{1}{K}\frac{1}{L}\sum_{k=1}^{K}\sum_{l=1}^{L} 10\log_{10}\frac{\| A_{km(k,l)}(z)s_{m(k,l)}(t-\frac{D}{2})\|^2}{\| y_k^{(\mathrm{ICA}l)}(t) - A_{km(k,l)}(z)s_{m(k,l)}(t-\frac{D}{2})\|^2} \quad (13)$$

The SA indicates a degree of the similarity between the separated signals of the ICA and real SIMO-model-based signals.

4.2. Results and Discussion

Figure 4 shows the results of SA for different speaker combinations. The bars on the right of this figure correspond to the averaged results of each combination. In the averaged scores, the improvement of SA in SIMO-ICA is 9.5 dB compared with the conventional ICA. From these results, it is evident that the separated signals in the SIMO-ICA is obviously superior to that in the conventional ICA-based method.

In order to confirm that the separated signals of the SIMO-ICA maintain the spatial qualities of each sound source, we compared the shape of the real HRTFs with that of estimated HRTFs by SIMO-ICA. We obtain the real and estimated HRTFs by truncating the corresponding early reflection components (within 25 taps after the peak pulse) from the real impulse responses $A_{ij}(z)\delta(t)$ and the estimated impulse responses. The impulse responses estimated by the SIMO-ICA are expressed as

$$[h_k^{(\mathrm{ICA}l)}(t)]_{k1} = \sum_{n=1}^{D-1} \boldsymbol{w}_{(\mathrm{ICA}l)}(n)\boldsymbol{A}(z) \begin{bmatrix} \delta(t+D/2-n) \\ \delta(t+D/2-n) \end{bmatrix}, \quad (14)$$

where $\delta(t+D/2-n)$ is used to cancel the time delay of $D/2$ in $\boldsymbol{w}_{(\mathrm{ICA}l)}(n)$. Figure. 5 shows the shapes of these HRTFs in both the best (SA=20.2 dB) and worst (SA=9.6 dB) cases. From this figure, it is confirmed that shapes of the estimated HRTFs are quite similar to the real HRTFs in the best case. Thus, we can conclude that SIMO-ICA has the potential to decompose the mixed binaural signals into SIMO-model-based signals without loss of information about spatial qualities of each sound source.

5. CONCLUSION

We apply single-input multiple-output (SIMO)-model-based ICA to the blind source separation problem of the binaural sounds. SIMO-ICA is the extended ICA algorithm for separating the mixed signals, not into monaural source signals but into SIMO-model-based signals of independent sources as they are at the microphones. In order to evaluate its effectiveness, separation experiments of binaural sounds recorded using HATS are carried out. The experimental results reveal that the performance of the proposed SIMO-ICA is superior to that of the conventional ICA-based method, and the separated signals of SIMO-ICA maintain the spatial qualities of each binaural sound source.

6. ACKNOWLEDGMENT

The authors are grateful to Dr. Shoji Makino, Ms. Shoko Araki, and Dr. Hiroshi Sawada of NTT Co., Ltd. for their discussions. This research is partly supported by Core Research for Evolutional Science and Technology program "Advanced Media Technology for Everyday Living" of Japan Science and Technology Agency.

7. REFERENCES

[1] P. Comon, "Independent component analysis, a new concept?," *Signal Processing,* vol.36, pp.287–314, 1994.

[2] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing,* vol.22, pp.21–34, 1998.

[3] H. Saruwatari, T. Kawamura, K. Shikano, "Blind source separation for speech based on fast-convergence algorithm with ICA and beamforming," *Proc. Eurospeech2001,* pp.2603–2606, Sept. 2001.

[4] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications,* John Wiley & sons, Ltd, West Sussex, 2002.

[5] J. Blauert, *Spatial Hearing (revised edition),* Cambridge, MA: The MIT Press, 1997.

[6] Y. Tatekura, H. Saruwatari, K. Shikano, "Sound reproduction system including adaptive compensation of temperature fluctuation effect for broad-band sound control," *IEICE Trans. Fundamentals,* vol.E85–A, no.8, pp.1851–1860, Aug. 2002.

[7] T. Takatani, T. Nishikawa, H.Saruwatari, K. Shikano, "High-fi delity blind source separation of acoustic signals using SIMO-model-based ICA with information-geometric learning," *Proc. IWAENC2003,* pp.251–254, 2003.

[8] S. Choi, S. Amari, A. Cichocki, R. Liu, "Natural gradient learning with a nonholonomic constraint for blind deconvolution of multiple channels," *Proc. International Workshop on ICA and BSS,* pp.371–376, 1999.

[9] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," *Proc. International Conf. on ICA and BSS,* pp.722–727, Dec. 2001.

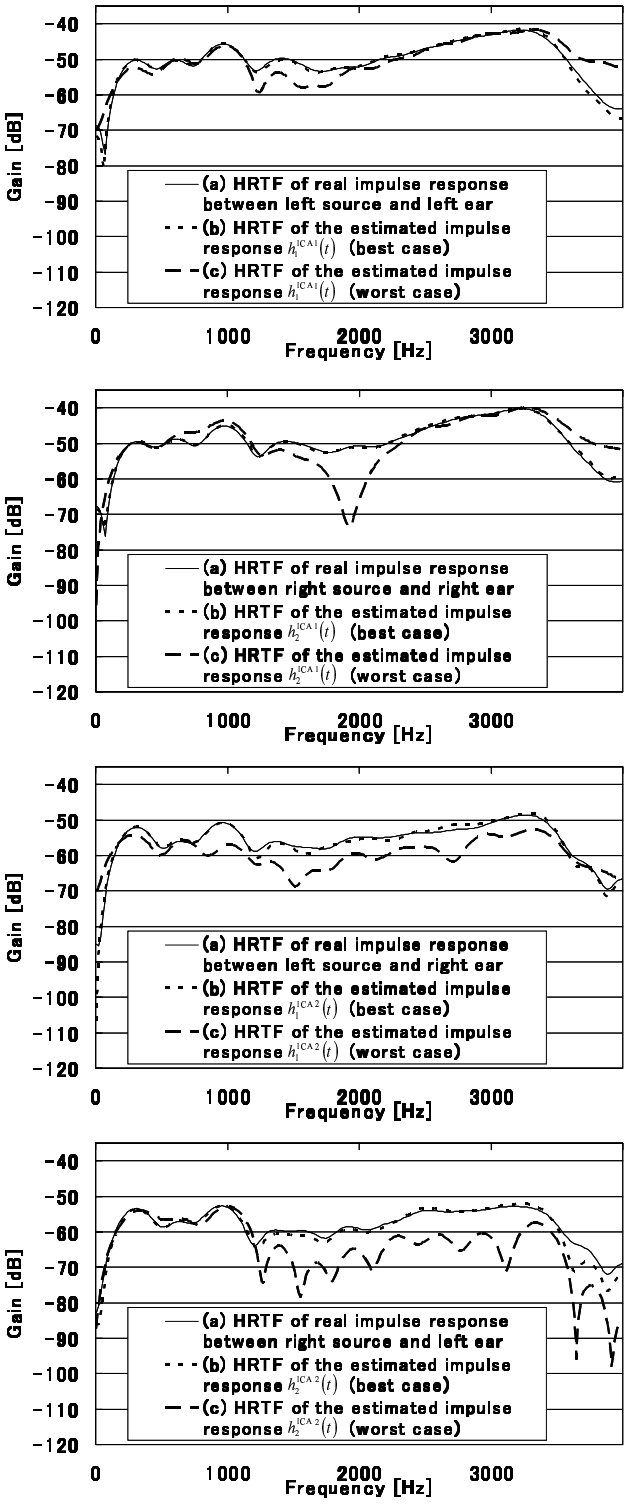

Fig. 5. Head-related transfer functions of (a) the real impulse responses $A_{ij}(z)\delta(t)$ and (b), (c) the impulse responses $h_k^{(\mathrm{ICA}l)}(t)$ estimated by SIMO-ICA in the best and worst cases.