Proceedings of 20th International Congress on Acoustics, ICA 2010

23–27 August 2010, Sydney, Australia

Noise robust semi-adaptive sound reproduction system based on semi-BSS

Yosuke Tatekura and Norihiro Yoshida

Faculty of Engineering, Shizuoka University
3-5-1 Johoku, Naka-ku, Hamamatsu, Shizuoka, 432-8561, Japan

PACS: 43.60.Bf, 43.60.Dh, 43.60.Pt

ABSTRACT

This paper proposes a reference signal extraction method for a semi-adaptive sound reproduction system based on semi-blind source separation (semi-BSS) under a noisy environment. Since the fluctuation of room transfer functions degrades the reproduced sound, we have proposed a semi-adaptive sound reproduction system that updates the inverse filters. However, in a noisy environment, it is difficult to observe only the reference signal. Therefore, we introduce semi-BSS based on frequency domain independent component analysis (FDICA) to the semi-adaptive sound reproduction system for extraction of the reproduced signal from the noisy observed signal. First, we obtain the noise signal from the noisy observed signal by semi-BSS. Next, the estimated reference signal is obtained by subtracting the noise signal from the observed signal. In a simulation using real environmental data, the proposed method can extract the reference signal with a high signal-to-deviation ratio (SDR) when RTFs were changed by temperature fluctuation.

INTRODUCTION

In order to realize a sound reproduction system with several loudspeakers, it is important to design inverse filters which cancel the effects of room transfer functions (RTFs) (Bauck and Cooper 1996). The RTFs vary depending on environmental variations (such as variations of speed of sound due to temperature fluctuations, change of reflection conditions due to changes in indoor items, etc.) and are not time invariant. Therefore, the reproduction accuracy is deteriorated by environmental variations in a sound reproduction system using a fixed inverse filter coefficients. Also, if unstable inverse filters enlarging the original signal are used, variation of the RTFs cause deterioration of sound quality, it is necessary to either re-estimate the RTFs after variations or to adaptively relax the inverse filters. Therefore, in sound reproduction systems that use fixed inverse filters, reproduction accuracy is degraded by fluctuations in the environmental. As an adaptive design procedure for the inverse filters, a method has been proposed for updating the inverse filter coefficients by means of reference microphones set up at several control points (Elliott et al. 1987, Nelson et al. 1992). However, since the reference microphones must be placed near the ears of the listener, hearing is significantly impaired.

We previously proposed a semi-adaptive sound reproduction system that compensates for environmental fluctuations, such as temperature fluctuations (Tatekura et al. 2002, Yai et al. 2008), and an inverse filter relaxation algorithm (Tatekura et al. 2005) in order to maintain the quality of the reproduced sound. In this system, one monitoring microphone can be placed at a location that does not restrict the listener, and the inverse filters can be updated by the signal observed at the monitoring microphone as a reference signal. However, since the environment normally contains several noise sources, it is difficult to observe only the reproduced signal by the conventional system.

To resolve this problem, we herein propose a method for observing only the reproduced signal in a noisy environment using the semi-adaptive sound reproduction system. In the proposed method, we introduce semi-blind source separation (semi-BSS) (Even et al. 2008) to the semi-adaptive sound reproduction system for extraction of the reproduced signal from the noisy observed signal.

SEMI-ADAPTIVE SOUND REPRODUCTION SYSTEM

Sound reproduction system and inverse filters

In this section, we describe the design method of the inverse filter in the frequency domain for sound reproduction system. In the following, we assume a multichannel sound reproduction system with M secondary sound sources L_m $(m = 1,2,\cdots,M)$ and N control points C_n $(n = 1,2,\cdots,N)$. We define the matrix representing the RTF, the inverse filter, the original sound source signal, and the reproduced sound be $\mathbf{G}(\omega)$, $\mathbf{H}(\omega)$, $\mathbf{D}(\omega)$, and $\mathbf{X}(\omega)$, respectively.

The matrices can be expressed as follows:

$$\mathbf{G}(\omega) = \begin{bmatrix} G_{11}(\omega) & G_{12}(\omega) & \dots & G_{1M}(\omega) \\ G_{21}(\omega) & G_{22}(\omega) & \dots & G_{2M}(\omega) \\ \vdots & \vdots & \ddots & \vdots \\ G_{N1}(\omega) & G_{N2}(\omega) & \dots & G_{NM}(\omega) \end{bmatrix} \tag{1}$$

$$\mathbf{H}(\omega) = \begin{bmatrix} H_{11}(\omega) & H_{12}(\omega) & \dots & H_{1N}(\omega) \\ H_{21}(\omega) & H_{22}(\omega) & \dots & H_{2N}(\omega) \\ \vdots & \vdots & \ddots & \vdots \\ H_{M1}(\omega) & H_{M2}(\omega) & \dots & H_{MN}(\omega) \end{bmatrix} \tag{2}$$

$$\mathbf{D}(\omega) = [D_1(\omega), D_2(\omega), \cdots, D_N(\omega)]^T \tag{3}$$

$$\mathbf{X}(\omega) = [X_1(\omega), X_2(\omega), \cdots, X_N(\omega)]^T, \tag{4}$$

where ω denotes the frequency, $G_{ji}(\omega)$ is the RTF and $H_{ij}(\omega)$ is the inverse filter coefficient. $i\ (= 1,2,\cdots,M)$ is the order of the secondary sound source and $j\ (= 1,2,\cdots,N)$ is the order of the control point. $D_j(\omega)$ is the original sound reproduced at control point j and $X_j(\omega)$ is the reproduced sound at control point j.

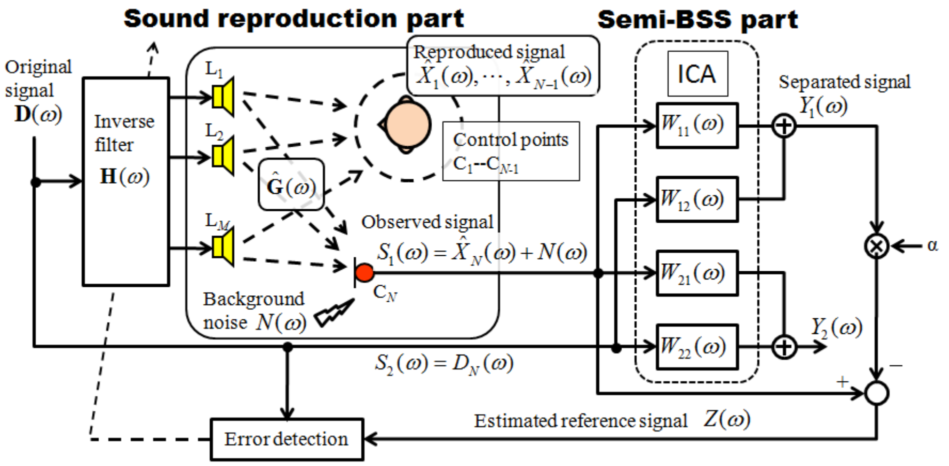
Sound reproduction part
Semi-BSS part
Original signal $\mathbf{D}(\omega)$
Inverse filter $\mathbf{H}(\omega)$
L_1
L_2
L_M
$\hat{\mathbf{G}}(\omega)$
Reproduced signal $\hat{X}_1(\omega), \cdots, \hat{X}_{N-1}(\omega)$
Control points C_1--C_{N-1}
Observed signal $S_1(\omega) = \hat{X}_N(\omega) + N(\omega)$
C_N
Background noise $N(\omega)$
$S_2(\omega) = D_N(\omega)$
ICA
$W_{11}(\omega)$
$W_{12}(\omega)$
$W_{21}(\omega)$
$W_{22}(\omega)$
Separated signal $Y_1(\omega)$
$Y_2(\omega)$
α
−
+
Estimated reference signal $Z(\omega)$
Error detection

Figure 1: Extraction of the reproduced signal from the noisy observed signal by semi-BSS in the semi-adaptive sound reproduction system.

The reproduced signal $\mathbf{X}(\omega)$ can be expressed as follows in terms of the matrices given above:

$$\mathbf{X}(\omega) = \mathbf{G}(\omega)\mathbf{H}(\omega)\mathbf{D}(\omega). \tag{5}$$

Since our objective is to achieve control such that $\mathbf{D}(\omega) = \mathbf{X}(\omega)$ in sound reproduction, the inverse filter $\mathbf{H}(\omega)$ can be obtained by deriving the inverse matrix of the room transfer function $\mathbf{G}(\omega)$. That is, the inverse filter design method is reduced to solving the following linear equation:

$$\mathbf{G}(\omega)\mathbf{H}(\omega) = \mathbf{I}_N, \tag{6}$$

where $\mathbf{I}_N$ is the $N \times N$ identity matrix. The inverse filter $\mathbf{H}(\omega)$ can be derived as the generalized inverse matrix of $\mathbf{G}(\omega)$ in the case of $M > N$. Since the solution becomes underdetermined if there is no rank reduction, the generalized Moore-Penrose (MP) inverse matrix with the least norm solution (LNS) (Kaminuma et al. 1999) is used. In the following, the generalized MP inverse matrix of $\mathbf{G}(\omega)$ is expressed as $\mathbf{G}^{\dagger}$. In order to derive the generalized MP inverse matrix, the singular value decomposition (SVD) of $\mathbf{G}(\omega)$ is carried out as follows:

$$\mathbf{G}(\omega) = \mathbf{U}(\omega) \cdot \left[\mathbf{P}_N(\omega),\ \mathbf{O}_{N,M-N}\right] \cdot \mathbf{V}^{\mathrm{H}}(\omega) \tag{7}$$

$$\mathbf{P}_N(\omega) \equiv \mathrm{diag}\left[\mu_1(\omega), \ldots, \mu_N(\omega)\right], \tag{8}$$

where $\mathbf{U}(\omega)$ is $N \times N$ orthogonal matrix, $\mathbf{V}(\omega)$ is $M \times M$ orthogonal matrix, $\mathbf{V}^{\mathrm{H}}(\omega)$ is the Hermitian transposed matrix of $\mathbf{V}(\omega)$, and $\mathbf{O}_{N,M-N}$ indicates the $N \times (M-N)$ null matrix. Also, $\mu_k(\omega)$ $(k = 1, \ldots, N,\ \mu_k(\omega) \geq \mu_{k+1}(\omega))$ denotes the singular values. By using Eq.(7), the generalized MP inverse matrix of $\mathbf{G}(\omega)$, $\mathbf{G}^{\dagger}(\omega)$, can be given by

$$\mathbf{G}^{\dagger}(\omega) = \mathbf{V}(\omega) \cdot \begin{bmatrix} \mathbf{P}_N^{-}(\omega) \\ \mathbf{O}_{M-N,N} \end{bmatrix} \cdot \mathbf{U}^{\mathrm{H}}(\omega), \tag{9}$$

where

$$\mathbf{P}_N^{-}(\omega) \equiv \mathrm{diag}\left[\xi_1(\omega), \ldots, \xi_N(\omega)\right] \tag{10}$$

$$\xi_k(\omega) = \begin{cases} \frac{1}{\mu_k(\omega)} & \mu_k(\omega) \neq 0 \\ 0 & \text{otherwise} \end{cases}. \tag{11}$$

By computing the inverse matrix $\mathbf{G}^{\dagger}(\omega)$ for each frequency, the inverse filter $\mathbf{H}(\omega)$ can be designed.

Semi-adaptation and its problem

For the cases in which the RTFs do not fluctuate, the observed signal $\mathbf{X}(\omega)$ can be written as Eq. (5). However, the RTFs are not time invariant and vary with fluctuations in the environment. If $\hat{\mathbf{G}}(\omega)$ represents the RTFs obtained after the fluctuation of $\mathbf{G}(\omega)$, then $\hat{\mathbf{G}}(\omega)$ is given by

$$\hat{\mathbf{G}}(\omega) = \mathbf{G}(\omega) + \Delta\mathbf{G}(\omega), \tag{12}$$

where

$$\Delta\mathbf{G}(\omega) = \begin{bmatrix} \Delta G_{11}(\omega) & \Delta G_{12}(\omega) & \ldots & \Delta G_{1M}(\omega) \\ \Delta G_{21}(\omega) & \Delta G_{22}(\omega) & \ldots & \Delta G_{2M}(\omega) \\ \vdots & \vdots & \ddots & \vdots \\ \Delta G_{N1}(\omega) & \Delta G_{N2}(\omega) & \ldots & \Delta G_{NM}(\omega) \end{bmatrix} \tag{13}$$

expresses the difference between the original RTFs $\mathbf{G}(\omega)$ and the fluctuated RTFs $\hat{\mathbf{G}}(\omega)$. Here, the observed signal after fluctuation $\hat{\mathbf{X}}(\omega)$ can be expressed as

$$\begin{aligned} \hat{\mathbf{X}}(\omega) &= \hat{\mathbf{G}}(\omega)\mathbf{H}(\omega)\mathbf{D}(\omega) \\ &= \mathbf{D}(\omega) + \Delta\mathbf{G}(\omega)\mathbf{H}(\omega)\mathbf{D}(\omega). \end{aligned} \tag{14}$$

The left-hand side of Fig. 1 shows an overview of a semi-adaptive sound reproduction system. Let the Nth control point be the control point for signal observation. The monitoring microphone C_N is set apart from the listener so as to promote agreeable listening. The observed signal after fluctuation $\hat{X}_N(\omega)$ observed at the monitoring microphone C_N is also given by

$$\hat{X}_N(\omega) = D_N(\omega) + \Delta\mathbf{G}_N(\omega)\mathbf{H}(\omega)\mathbf{D}(\omega), \tag{15}$$

where

$$\Delta\mathbf{G}_N(\omega) = \left[\Delta G_{N1}(\omega),\ \Delta G_{N2}(\omega),\ \cdots,\ \Delta G_{NM}(\omega)\right] \tag{16}$$

represents the RTFs from each loudspeaker to the monitoring microphone. Using the observed signal $\hat{X}_N(\omega)$ as the reference signal, all of the inverse filters can be updated.

However, these semi-adaptive algorithms are assumed to detect only the reproduced sound from the system by the monitoring microphone. If there are any noise sources inside or outside the reproduced room, the monitoring microphone detects $\hat{X}_N(\omega) + N(\omega)$ rather than $\hat{X}_N(\omega)$, where $N(\omega)$ is the noise signal. As such, when several background noises, such as the voice of the user, are also observed, the algorithms do not necessarily work correctly. Therefore, we must extract only the reproduced signal $\hat{X}_N(\omega)$ from $\hat{X}_N(\omega) + N(\omega)$.

REFERENCE SIGNAL EXTRACTION UNDER NOISY ENVIRONMENT

Semi-blind source separation

Blind source separation (BSS) based on frequency domain independent component analysis (FDICA) is a technique for estimating an original sound source based solely on the mixed signals observed at each microphone (Murata and Ikeda 1998). This technique does not necessarily require prior information such as the direction of the target sound source or speech break. If one of original sound sources is known, then conventional FDICA-based BSS can be developed into semi-BSS (Even et al. 2008). In the proposed semi-adaptive sound reproduction system, we introduce the semi-BSS method to obtain the reference signal from a noisy observed signal as shown in Fig.1.

The separated signal of semi-BSS, i.e.,

$$\mathbf{Y}(\omega) = [Y_1(\omega), Y_2(\omega)]^T, \tag{17}$$

can be written in terms of the 2×2 separation matrix $\mathbf{W}(\omega)$ and the input signal

$$\mathbf{S}(\omega) = [S_1(\omega), S_2(\omega)]^T \tag{18}$$

as follows:

$$\mathbf{Y}(\omega) = \mathbf{W}(\omega)\mathbf{S}(\omega). \tag{19}$$

In semi-BSS, the component $S_2(\omega)$ of $\mathbf{S}(\omega)$ is assumed to be a known signal, and so we can set $S_2(\omega) = D_N(\omega)$. The input signal $S_1(\omega)$ is set to be the observed signal at the monitoring microphone, which can be written as

$$S_1(\omega) = \hat{X}_N(\omega) + N(\omega). \tag{20}$$

We can write $\mathbf{W}(\omega)$ in semi-BSS as follows:

$$\mathbf{W}(\omega) = \begin{bmatrix} W_{11}(\omega) & W_{12}(\omega) \\ 0 & W_{22}(\omega) \end{bmatrix}. \tag{21}$$

Here, $\mathbf{W}(\omega)$ is optimized with the following iterative update procedure:

$$\mathbf{W}(\omega) \leftarrow \mathbf{W}(\omega) + \mu \left(\mathbf{I} - E[\varphi(\mathbf{Y}(\omega))\mathbf{Y}(\omega)^H] \right) \mathbf{W}(\omega), \tag{22}$$

where μ is the step-size parameter, $\mathbf{I}$ is the 2×2 identity matrix, and $\varphi(\cdot)$ is a nonlinear vector function.

Note that, in actual BSS, the short-time analysis of the input signals is performed using DFT in a frame-by-frame manner.

Reference signal extraction from observed signal

The reproduced signal picked up by the monitoring microphone $X_N(\omega)$ is expected to be equivalent to the original sound source signal $D_N(\omega)$ in the absence of environmental fluctuation. Although the RTFs fluctuated as shown in Eq. (12), the original signal $D_N(\omega)$ and the fluctuated reproduced signal $\hat{X}_N(\omega)$ are highly similarity. Therefore, $\hat{X}_N(\omega)$ can be separated with high accuracy from noisy observed signal $S_1(\omega)$ using $D_N(\omega)$ by semi-BSS. However, semi-BSS extracts an unknown signal from the observed signal using a known signal. In other words, the noise signal, and not the reproduced signal, can be obtained directly. The goal is to obtain $\hat{X}_N(\omega)$ from noisy observed signal $S_1(\omega)$.

Therefore, we first obtain the noise signal $N(\omega)$ from $S_1(\omega)$ by semi-BSS. Next, the estimated reference signal $Z(\omega)$, which is assumed to be $\hat{X}_N(\omega)$, is obtained by subtracting the noise signal from the observed signal, as follows:

$$Z(\omega) = S_1(\omega) - \alpha Y_1(\omega), \tag{23}$$

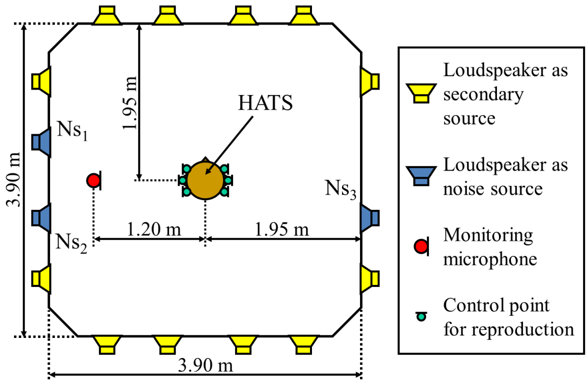
Loudspeaker as secondary source
Loudspeaker as noise source
Monitoring microphone
Control point for reproduction
HATS
Ns$_1$
Ns$_2$
Ns$_3$
1.95 m
3.90 m
1.20 m
1.95 m
3.90 m

Figure 2: Layout of acoustic experiment room.

where α represents the scaling parameter. An optimized α, $\hat{\alpha}$, is obtained in order to minimize the spectrum distortion (SD) between $Z(\omega)$ and the original source signal $D_N(\omega)$:

$$\begin{aligned} \hat{\alpha} &= \operatorname*{argmin}_{\alpha} \sqrt{\frac{1}{N}\sum_{\omega}\left(20\log\frac{|Z(\omega)|}{|D_N(\omega)|}\right)^2} \\ &= \operatorname*{argmin}_{\alpha} \sqrt{\frac{1}{N}\sum_{\omega}\left(20\log\frac{|S_1(\omega) - \alpha Y_1(\omega)|}{|D_N(\omega)|}\right)^2}. \end{aligned} \tag{24}$$

NUMERICAL EVALUATION

Experimental setup

To investigate the effect of the proposed method, a numerical simulation with real environmental data was carried out. A sound reproduction system having 12 loudspeakers and 7 control points (6 for listening and 1 for monitoring) was set up in the experimental room, in which the reverberation is approximately 0.14 seconds. The arrangement is shown in Fig. 2.

The impulse responses of the RTFs used in this simulation are obtained by a time stretched pulse signal (Suzuki et al. 1995) of 65,536 points, where the sampling frequency is 48,000 Hz, the quantization is 16 bits, and averaging is performed four times. In this simulation, these impulse responses are downsampled from 48,000 Hz to 8,000 Hz. The fluctuation of the RTFs is assumed to be caused by temperature variation in the room, and the impulse responses were measured at several temperatures. The inverse filters with 16384 points are designed by the MP generalized inverse matrix of $\mathbf{G}(\omega)$ under the initial temperature condition.

A music signal of 9 seconds is used as the original signal for reproduction, and different music signal is used as the original signal for monitoring. Three signals, namely, a telephone ring tone, a clock alarm, and a speech signal were used for background noise.

As the analysis conditions of semi-BSS, the frame length was 256 ms, and the frame shift was 64 ms. In Eq. (22), the tanh function is used as the nonlinear function. The iteration times and step size for which the extraction accuracy is maximum are set by preliminary experiments at each input SNR. The input-SNR is defined as follows:

$$\text{input-SNR [dB]} = 10\log_{10}\frac{\sum_n |\hat{x}_N(n)|^2}{\sum_n |n(n)|^2}, \tag{25}$$

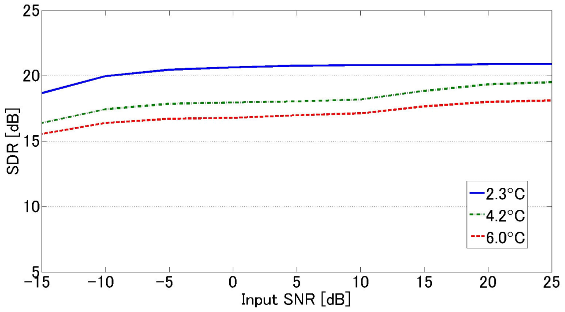

Figure 3: Signal to deviation ratio of the extracted signal for different input-SNR conditions for different temperature fluctuations.

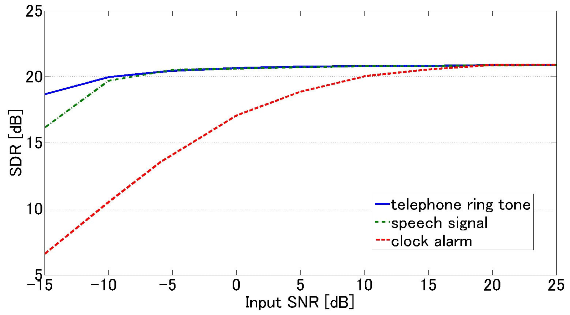

Figure 4: Signal to deviation ratio of the extracted signal for different input-SNR conditions for different background noise conditions.

where $\hat{x}_N(n)$ and $n(n)$ are time domain representations of $\hat{X}_N(\omega)$ and $N(\omega)$, respectively.

Experimental results

Figure 3 shows the result for extraction accuracy under different temperature fluctuation conditions using the telephone ring tone as the background noise signal. The signal to deviation ratio (SDR) is used to evaluate the extraction accuracy, which can be computed as follows:

$$\mathrm{SDR}\ [\mathrm{dB}] = 10\log_{10}\frac{\sum_n |z(n)|^2}{\sum_n |z(n)-\hat{x}_N(n)|^2}, \qquad (26)$$

where $z(n)$ is time domain representations of $Z(\omega)$. The loudspeaker Ns$_1$ in Fig.2 was used as noise source in any case. This result shows that high extraction accuracies are maintained under all of the temperature fluctuation conditions.

Figure 4 shows the results for extraction accuracy for different background noises where the temperature fluctuation is 2.3°C. The loudspeaker Ns$_1$ in Fig.2 was used as noise source in this case also. The figure indicates large differences in the extraction accuracy depending on the type of background noise at lower input-SNRs. However, SDR of over 15 dB can be obtained under any condition using the proposed method considered herein when the input-SNR exceeds 0 dB.

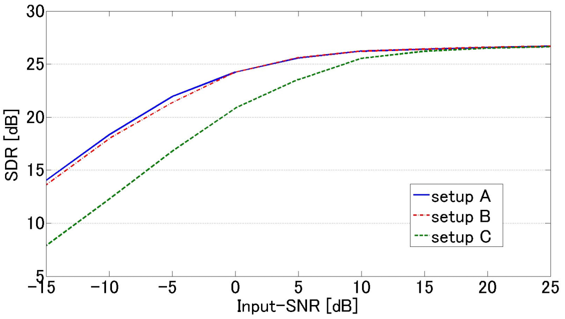

Figure 5: Signal to deviation ratio of the extracted signal for different noise sources arrangement for two background noise conditions.

Table 1: Combinations of the noise sources and loudspeakers

setup A	Ns$_1$: telephone ring tone Ns$_1$: speech signal
setup B	Ns$_1$: telephone ring tone Ns$_2$: speech signal
setup C	Ns$_1$: telephone ring tone Ns$_3$: speech signal

Figure 5 shows the results for extraction accuracy under two background noises conditions where the temperature fluctuation is 2.3°C. The combinations of the background noise signals and loudspeakers as noise sources are listed in table 1. In this case, another speech signal of 9 seconds is used as the original signal for reproduction. This figure reveals that each combination shows almost similar tendency though SDR of setup C is a little low at lower input-SNRs. As a result, we can find that the performance of the proposed method is strongly influenced by the input-SNR, and is not influenced so much by the number of background noises.

CONCLUSION

We have proposed an algorithm to extract a reference signal from a noisy observed signal by semi-BSS for achievement of noise robust semi-adaptive sound reproduction system. In simulations using real environmental data, the proposed method was demonstrated to extract the reference signal with a high SDR. In addition, the results reveal that the proposed method is robust against temperature fluctuation of the reproduced room and various types of noise. In future studies, we intend to apply the proposed method to an actual semi-adaptive sound reproduction system and evaluate its efficiency by subjective evaluation.

REFERENCES

J. Bauck and D. H. Cooper. Generalized transaural stereo and applications. *J. Audio Eng. Soc.*, 44(9):683–705, 1996.

S. J. Elliott, I. M. Stothers, and P. A. Nelson. A multiple error lms algorithm and its application to the active control of sound and vibration. *IEEE Trans. ASSP*, 35(10):1423–1434, 1987.

J. Even, H. Saruwatari, and K. Shikano. Frequency domain semi-blind signal separation: Application to the rejection of internal noises. *Proc. ICASSP2008*, pages 157–160, 2008.

A. Kaminuma, S. Ise, and K. Shikano. A method of designing inverse system for multi-channel sound reproduction system using least-norm-solution. *Proc. Active99*, 2:863–874, 1999.

N. Murata and S. Ikeda. An on-line algorithm for blind source separation on speech signals. *Proc. NOLTA98*, pages 923–926, 1998.

P. A. Nelson, H. Hamada, and S. J. Elliott. Adaptive inverse filters for stereophonic sound reproduction. *IEEE Trans. Signal Processing*, 40(7):1621–1632, 1992.

Y. Suzuki, F. Asano, H.-Y. Kim, and T. Sone. An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses. *J. Acoust. Soc. Am.*, 97(2):1119–1123, 1995.

Y. Tatekura, H. Saruwatari, and K. Shikano. Sound reproduction system including adaptive compensation of temperature fluctuation effect for broad-band sound control. *IEICE Trans. Fundamentals*, E85-A(8):1851–1860, 2002.

Y. Tatekura, S. Urata, H. Saruwatari, and K. Shikano. On-line relaxation algorithm applicable to acoustic fluctuation for inverse filter in multichannel sound reproduction system. *IEICE Trans. Fundamentals*, E88-A(7):1747–1756, 2005.

Y. Yai, S. Miyabe, H. Saruwatari, K. Shikano, and Y. Tatekura. Rapid compensation of temperature fluctuation effect for multichannel sound field reproduction system. *IEICE Trans. Fundamentals*, E91-A(6):1329–1336, 2008.