3D Sound Field Recording and Reproducing System including Sound Source Orientation

Toshimasa Suzuki*, Hirofumi Nakajima†, Hideo Tsuru‡, Takayuki Arai* and Kazuhiro Nakadai†
*Graduate School of Science and Technology, Sophia University
7-1 Kioi-cho, Chiyoda-ku, Tokyo, 102-8554, Japan
†Honda Research Institute Japan Co., Ltd.
8-1 Honcho, Wako-shi, Saitama, 351-0114, Japan
‡Nittobo Acoustic Engineering Co., Ltd.
1-21-10 Midori, Sumida-ku, Tokyo 130-0021, Japan

***Abstract*—To achieve ultra-realistic communications (URC), a three-dimensional (3D) sound field recording and reproducing system using spatial sound features is promising. It utilizes sound source positions and original source signals as features. On the other hand, sound source orientations (such as speaker's orientation) have been neglected, though they are necessary for high-performance URC. In this paper, we propose a 3D sound field recording and reproducing system that includes sound source orientation. For the recording part, we applied and evaluated a sound source orientation estimation method based on orientation-extended beamforming. This method requires transfer functions for all possible source positions and orientations, which should be either measured or calculated. Measured transfer functions have high accuracy because all characteristics, such as reverberations and diffractions, are taken into account. Calculated transfer functions are obtained for any environment without measurements. We performed experiments to evaluate the orientation estimation method using transfer functions obtained by both practical measurements and calculations using acoustic simulation based on wave theory. The experimental results show that our estimation method has sufficient capability using both measured and calculated transfer functions. For the reproducing part, we evaluated impulse responses versus sound source orientation. The evaluation results revealed significant differences in both objective and subjective scores. This proves that our system achieves higher performance than conventional ones that do not utilize orientation features.**

I. Introduction

To achieve ultra-realistic communications (URC), people have studied recording and reproducing techniques for image and acoustic information [1], [2]. For acoustic information, a recording and reproducing system using multiple microphones (non-parametric method) is often used [3]. This system achieves URC, but must treat a large amount of data because of the multichannel acoustic waveform signals and does not have enough flexibility for various manipulations, such as position changes for individual sound sources. On the other hand, the parametric method records and reproduces a three-dimensional (3D) sound field using only its features and requires less data, so it is suitable for many applications [2]. The common features are the sound source positions and original source signals, which are obtained by sound source localization and separation methods, respectively. For this method, sound source orientation has not been used, even though sound source orientation is important for recording 3D sound field acoustic features with high accuracy. For example, in a multi-party situation, to understand whom a speaker is talking to, one requires the speaker's orientation. Similarly, for realistic reproduction of a musical performance using multiple instruments in a concert hall, instruments directivities are important. Consequently, a sound source orientation estimation method is essential to achieve URC. In this paper, we propose a parametric 3D sound field recording and reproducing system that includes sound source orientation.

II. System overview

An overview of our 3D sound field recording and reproducing system is shown in Fig. 1. The system consists of recording and reproducing parts.

For the recording part, we use a smart-room that has a microphone array. The multi-input signals from the sound sources (human and instruments) are recorded by the microphone array. The sound sources (for example, human and instruments) in this room are labeled $S_1, S_2, \cdots, S_n$. Acoustic waveform signals from the sound sources are recorded as multi-input signals by the microphone array. Using these signals, the sound source localization part estimates each sound source position (x, y, z for S_1 to S_n). After that, the estimated data and signals are used by the sound source separation and orientation estimation methods to estimate the original signal of each sound source ($s_1(t), s_2(t), \cdots, s_n(t)$) and each sound source orientation (θ, ϕ for S_1 to S_n). The sound source orientation estimation requires transfer functions for all possible combinations of the source positions and orientations. The transfer functions are stored as a database obtained by practical measurements or acoustic simulations in advance. For these processes, 3D sound field parametric data is estimated.

In the reproducing part, the 3D sound field features are applied to create a reproducing field. The field can be a real or virtual one. For a real field, a reproducing field and equipment such as microphones and loudspeakers are required. On the other hand, for a virtual field, a simulation engine is required. Location data and orientation data for S_1 are applied and manipulated for the reproducing field's sound source as a

978-1-4244-7820-0/10/$26.00 ©2010 IEEE

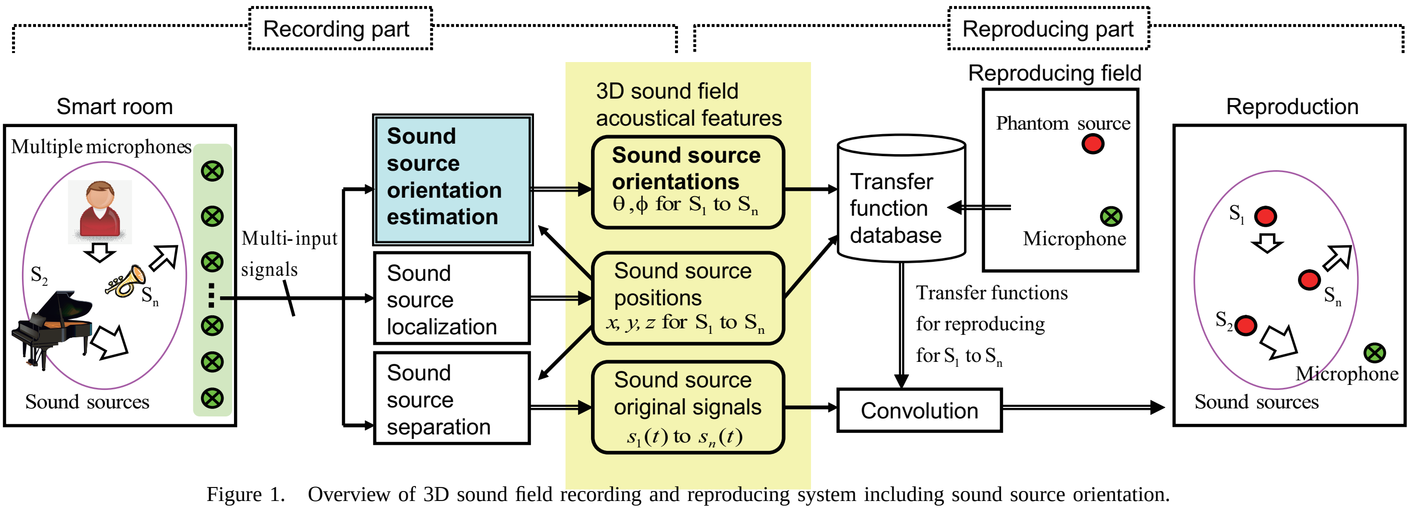

Figure 1. Overview of 3D sound field recording and reproducing system including sound source orientation.

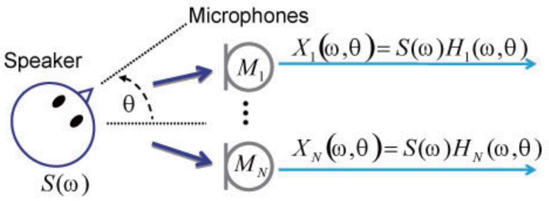
Microphones
Speaker
θ
$S(\omega)$
M_1
M_N
$X_1(\omega,\theta)=S(\omega)H_1(\omega,\theta)$
$X_N(\omega,\theta)=S(\omega)H_N(\omega,\theta)$

Figure 2. Model of wave propagation including sound source orientation.

phantom source for S_1. The transfer function for reproducing is obtained using the impulse response from the phantom source to the microphone. Then, the synthesized signal for S_1 is calculated from the convolution of the original signal ($s_1(t)$) and this transfer function. In a similar way, the synthesized signals for S_2 to S_n are calculated. As a result, the reproduced signal is obtained for the combination of all the synthesized signals.

Since sound source localization and separation are commonly used techniques [4], [5], we focus on the sound source orientation estimation in the following section.

III. Sound Source Orientation Estimation

Okamoto *et al.* proposed a method to estimate all-around directivity of a sound source, although this method is not applicable for sound source separation [6]. We use the orientation-extended beamforming method for sound source orientation estimation. Beamforming is one of the methods for sound source localization and applicable for sound source separation. Orientation-extended beamforming can estimate not only sound source locations but also sound source orientations by using transfer functions extended for sound source orientation [7]. This beamforming process is equivalent to pattern matching between the input signal and the transfer function database. Here, we show the algorithm of orientation-extended beamforming using database matching.

A. Transfer function extended for sound source orientation

A model of wave propagation including sound source orientation using a microphone array is shown in Fig. 2. $S(\omega)$ is the frequency response of a sound source (human speaker), where ω is the angular frequency. M_k is the k-th microphone ($k = 1, 2, \cdots, N$). $H_k(\omega, \theta)$ is the transfer function from the sound source orientated at $\theta°$ to the k-th microphone. We assume that the sound source position has been determined through a localization process and abbreviate the position variables in the following equations. The received signal X_k at microphone M_k is represented by

$$X_k(\omega) = S(\omega)H_k(\omega). \tag{1}$$

Transfer functions are represented by a vector:

$$\boldsymbol{h}(\omega,\theta) = [|H_1(\omega,\theta)|, \cdots, |H_N(\omega,\theta)|]^T, \tag{2}$$

where T is the transpose. In this equation, we take only amplitude components from $H_N(\omega, \theta)$ because phase components are easily affected by environmental changes in the high-frequency band, so this amplitude extraction avoids precision degradation.

B. Transfer function database

We make a transfer function database as a set of transfer function vectors. The transfer function vector from a sound source is calculated as

$$\boldsymbol{h}_0(\omega,\theta) = \frac{\boldsymbol{h}(\omega,\theta)}{\sqrt{\boldsymbol{h}(\omega,\theta)^H \boldsymbol{h}(\omega,\theta)}} = \frac{\boldsymbol{h}(\omega,\theta)}{|\boldsymbol{h}(\omega,\theta)|}, \tag{3}$$

where H is the transpose operation with complex conjugation. Therefore, $\boldsymbol{h}_0(\omega, \theta)$ does not include any features of the output equipment (loudspeaker frequency characteristics etc.).

C. Sound source orientation estimation using transfer function database

Input signals are represented by the received signal vector:

$$\boldsymbol{X}(\omega,\theta_s) = [|X_1(\omega,\theta_s)|, \cdots, |X_N(\omega,\theta_s)|]^T, \tag{4}$$

where θ_s is the sound source orientation (unknown data). $X_N(\omega, \theta_s)$ consists of only amplitude components. The normalized $\boldsymbol{X}(\omega, \theta_s)$ is calculated as

$$\boldsymbol{X}_0(\omega,\theta_s) = \frac{S(\omega)\boldsymbol{h}(\omega,\theta_s)}{|S(\omega)\boldsymbol{h}(\omega,\theta_s)|} = \frac{S(\omega)}{|S(\omega)|}\boldsymbol{h}_0(\omega,\theta_s). \tag{5}$$

We obtain the inner product between Eqs. (3) and (5) as

$$\begin{aligned} C_\omega(\omega,\theta) &= |\boldsymbol{h}_0(\omega,\theta)^H \boldsymbol{X}_0(\omega,\theta_s)| \\ &= \left| \frac{S(\omega)}{|S(\omega)|} \boldsymbol{h}_0(\omega,\theta_s) \right| \\ &= |\boldsymbol{h}_0(\omega,\theta)^H \boldsymbol{h}_0(\omega,\theta_s)|, \end{aligned} \tag{6}$$

where $C_\omega(\omega,\theta)$ represents the similarity between transfer functions for θ and θ_s.

The function of estimated orientation $C(\theta)$ without ω is calculated as

$$C(\theta) = \sum_\omega w(\omega,t) C_\omega(\omega,\theta), \tag{7}$$

where $w(\omega,t)$ is a weight function for masking non-speech and noise separation. To make the weight function, we used histogram-based recursive level estimation (HRLE) masking [8]. If $C(\theta)$ is maximum, θ is the estimated orientation $\hat{\theta}$.

D. Design of transfer function database

The transfer function database including all possible speaker positions and orientations should be prepared in advance. There are two ways to make transfer functions. One is to use practical measurement. Measured transfer functions have high accuracy because all characteristics, such as reverberations and diffractions, are taken into account. But such practical measurements are time-consuming and sometimes physically difficult. An automatic measurement system also has difficulties because it is only available in controlled environments and its cost is high. The other way is acoustic simulation. This makes it easy to change sound source positions and orientations and obtain the transfer functions. Therefore, if the environments are accurately simulated, any transfer functions can be obtained without any measurement. This makes our system more practical.

IV. Evaluation of Sound Source Orientation Estimation

We evaluated the precision of our sound source orientation estimation method. We calculated transfer functions using both practical measurement and acoustic simulation and evaluated the availability for sound source orientation estimation. Here, we describe these evaluations and show how to calculate highly accurate transfer functions using acoustic simulation that includes sound source directivity. The target accuracy was an estimation error of at most 45°.

A. Recording environment

The experimental room is shown in Fig. 3. The room size is 7.0 m × 4.0 m ×3.2 m. This room has 96 microphones embedded in the walls and kitchen shelf. The microphone positions are shown in Fig. 4. The reverberation time is approximately 230 ms. The sound source positions were set at the center of the room.

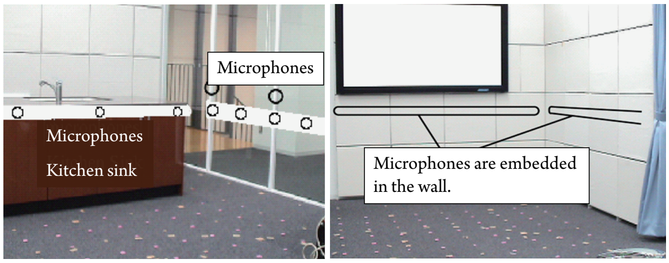

Figure 3. Experimental room.

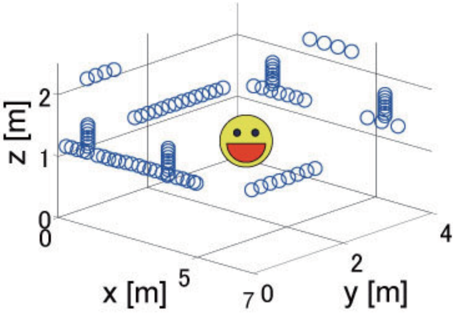

Figure 4. Microphone positions.

B. Calculation of transfer function database in practical measurement

We measured impulse responses at the 96 microphones using an automatic measuring system. The sound source was a loudspeaker (GENELEC 1029A) set at the room's center. The number of loudspeaker orientations was 8 (45° steps). As sound source signals, we used swept-sine signals having 2^{14} samples. The sampling frequency was 16 kHz.

C. Calculation of transfer function database by acoustic simulation

1) Simulation engine: To apply acoustic simulation to speaker orientation estimation, we need a simulation engine that has high numerical precision and can make directivity patterns. To meet these requirements, we used acoustic numerical calculation software COMFIDA version 2.03 (Nittobo Acoustic Engineering Co., Ltd.). This engine is based on a finite difference time domain (FDTD) method [9]. In acoustic simulation, calculation methods are classified into roughly two types: geometrical-based and wave-based methods. Geometrical-based methods simulate wave propagations as a trace, so they cannot simulate wave properties such as phase interference, diffraction, and eigenmodes. On the other hand, wave-based methods are based on wave acoustic theory, so they can simulate wave properties accurately. FDTD is a wave-based method. It can easily reflect physical values (density, velocity, etc.) and can reproduce room environment properties (size, reflection coefficient, etc.). Therefore, we chose an FDTD-based simulator. Moreover, the engine we chose uses a compact finite difference technique [10] to improve the numerical precision.

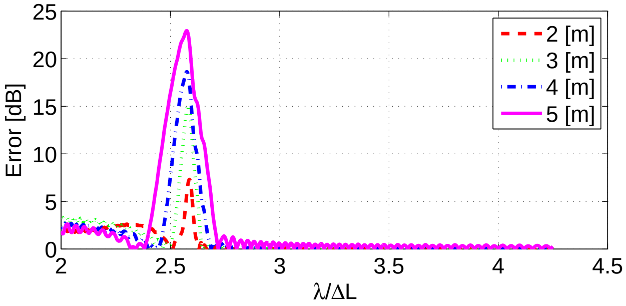

Figure 5. Error versus $\lambda/\Delta L$.

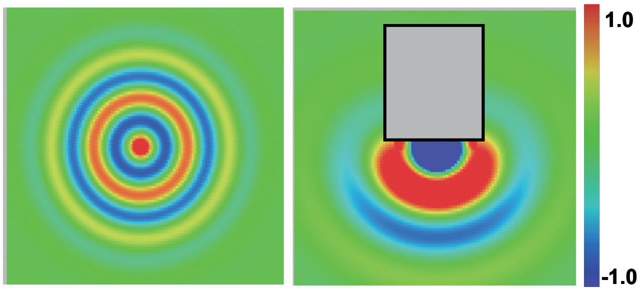

Figure 6. Sound pressure map.

2) Numerical precision evaluation: We evaluated the numerical precision of this engine. In the FDTD method, simulated space is digitized to a lattice. Numerical precision depends strongly on the ratio of the space lattice interval (ΔL) to the wavelength (λ). If $\lambda/\Delta L$ is not large enough, numerical errors in wave propagation grow in amplitude and phase components. These errors increase in proportion to the wave propagation distance. These effects are called 'numerical dispersion'. The relation between decay errors regarding propagation distance and $\lambda/\Delta L$ in this engine is shown in Fig. 5. These decay errors are represented by the difference in sound pressure levels obtained by calculation and by theoretical analysis. We evaluated the levels at points 2, 3, 4, and 5 m away from the sound source in comparison with 1 m away. For stable numerical precision, $\lambda/\Delta L$ should be set to more than 3.

3) Directivity pattern reproduction: In the FDTD method, we reproduced directivity patterns by modeling the shape of the sound source. A sound pressure map simulated by FDTD is shown in Fig. 6. In the left figure, the sound source was an omni-directional point source normally used in FDTD. To reproduce the directivity patterns, we used the FDTD method to combine a point source and the simplified source shape. The right figure is an example of directivity pattern reproduction. Although several studies on coping with sound source directivity have been reported, this method has the advantages of simplicity and applicability. It also has the advantage that the diffractions of every frequency are simulated precisely because FDTD is based on wave theory.

4) Loudspeaker directivity patterns: We simulated loudspeaker directivity patterns and evaluated them in the horizontal plane. The GENELEC 1029A loudspeaker is shown in Fig. 7 and the model designed for the simulation is shown in Fig. 8. This model is a combination of a rigid cube with a cylindrical hollow and a point source. This simple shape ignores the details of the loudspeaker's shape and properties, such as sound absorption coefficient and baffle vibrations. The parameters x, y, z, r, and d are given in Table I. Model 1 is the set of parameters closest to a practical loudspeaker shape. To evaluate the effects of parameter differences, we changed parameters r and d in Model 2 and x and y in Model 3. The loudspeaker directivity patterns at 1000 Hz obtained by simulation and practical measurement are shown in Fig. 9, where $\theta = 0°$ is the front of the loudspeaker. This figure shows that our method approximately reproduced the loudspeaker directivity patterns. Comparing practical measurement with Models 1, 2, and 3, we found that parameters r and d were less important than x and y.

Figure 7. GENELEC 1029A.

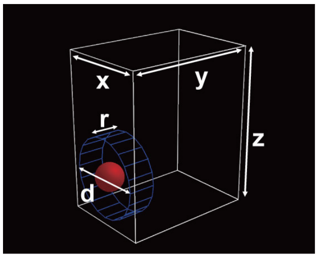

Figure 8. Loudspeaker model.

TABLE I
LOUDSPEAKER MODEL PARAMETERS.

Size	x, y, z, r, d [m]
Model 1	0.14, 0.20, 0.24, 0.02, 0.12
Model 2	0.14, 0.20, 0.24, 0.04, 0.08
Model 3	0.22, 0.10, 0.24, 0.02, 0.12

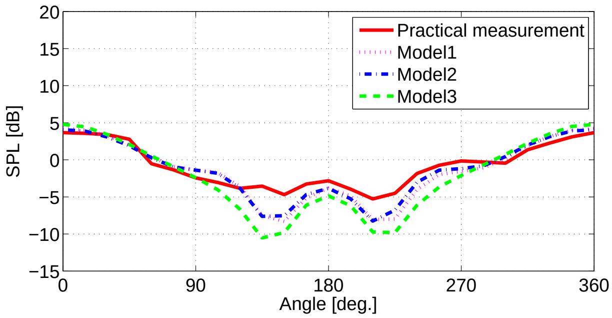

Figure 9. Directivity patterns at 1000 [Hz].

5) Calculation of transfer functions: We simulated impulse responses that had been measured in practice. The simulated experimental room is shown in Fig. 10. The many small spheres are evaluation points for simulating microphones. The loudspeaker was designed using the method described in the previous section. The modeling parameters were set to [x, y, z, r, d] = [0.14, 0.20, 0.24, 0, 0] because r and d were assumed

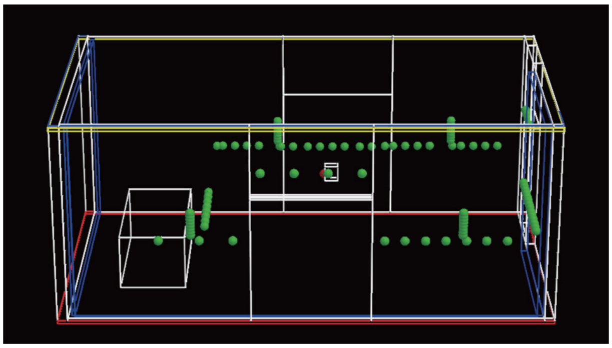

Figure 10. Simulated experimental room.

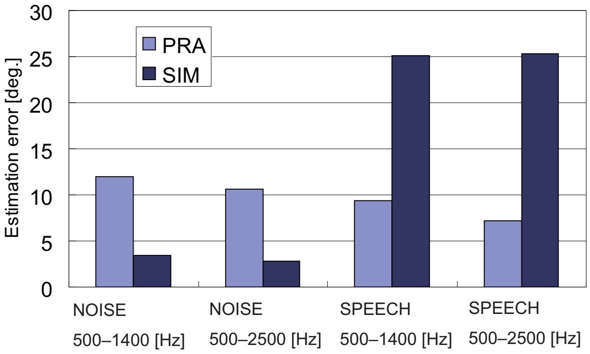

Figure 11. Estimation errors.

to be ineffective for directivity pattern representation. To reduce the amount of working memory required, we digitized the experimental room to a nonuniform lattice. Around the loudspeaker model and evaluation points, ΔL was 0.02 m; elsewhere, it was 0.08 m. The sound source produced pulse signals with flat frequency characteristics from 500 to 2500 Hz. The sampling frequency was 48 kHz. Considering numerical dispersion, the calculations were stable in the frequency band from 500 to 1400 Hz.

D. Sound source signal

We used speech and noise signals as the sound source signals. The speech signals were recorded at the center of the experimental room. A female speaker uttered “a, i, u, e, o”. The noise signals were calculated by convolution using impulse responses obtained by both practical measurement and acoustic simulation. For all signals, the number of sound source orientations was 4 (90° steps). All input signals were transformed into the frequency domain by fast Fourier transform (FFT) using 1024 points.

E. Results

The estimation errors for speaker orientations are shown in Fig. 11. This figure shows two frequency band patterns: 500–1400 Hz and 500–2500 Hz. The sound source signals were noise and speech signals, denoted NOISE and SPEECH, respectively, and measured and simulated transfer functions, denoted PRA and SIM, respectively. Errors are shown as averages of all orientations. In this figure, the errors for estimation using convolution signals for the sound source signals are approximately 10°. In the case where speech was used for the input data, the errors for estimation by acoustic simulation and practical measurement are about 25° and about 10°, respectively. These results show that the estimation accuracy is sufficiently greater than the target accuracy.

V. Evaluation of sound source orientation for reproducing

We evaluated the impulse responses with sound source orientation for reproduction. For reproduction, we used a geometrical acoustic simulation engine CATT-acoustic ver. 8.0. As discussed previously, this method cannot simulate sound wave characteristics exactly. However, it is sufficient to simulate echo patterns and impulse responses roughly. Moreover, it easily can simulate sound propagation in various sound fields, so it is commonly used for reproducing 3D sound fields. The reproducing field was a small hall (Fig. 12). The hall size is approximately 8.0 m ×21.0 m ×8.0 m. The microphone position was set at the center of the hall. The sound source position was the center of the hall’s stage 5.0 m away from the sound source. The sound source directivity was the cardioid pattern. The sound source orientation could be manipulated. (When the sound source was oriented toward the microphone, its orientation was 0°.) We reproduced impulse responses with eight orientations (45° steps) and made objective and subjective evaluations. These impulse responses were recorded binaurally.

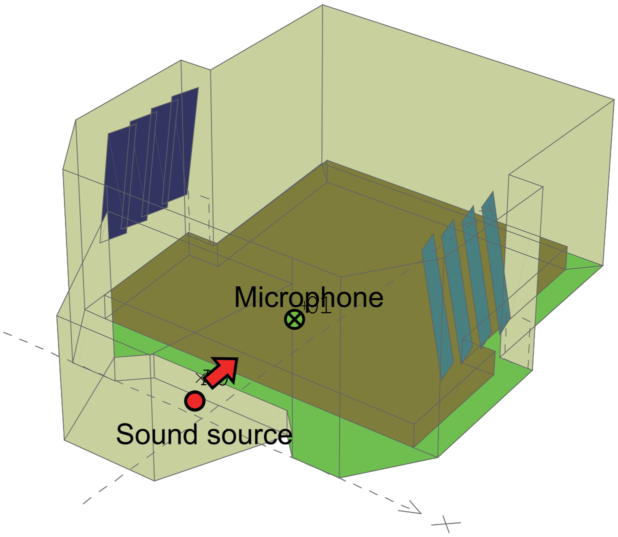

Figure 12. Reproducing room.

A. Objective evaluation

Impulse responses in the small hall are shown in Fig. 13. The temporal peaks and impulse response intensities were not the same.

The acoustic parameters for objective evaluation are given in Table II. Parameter D_{50} (Definition) is the ratio of early

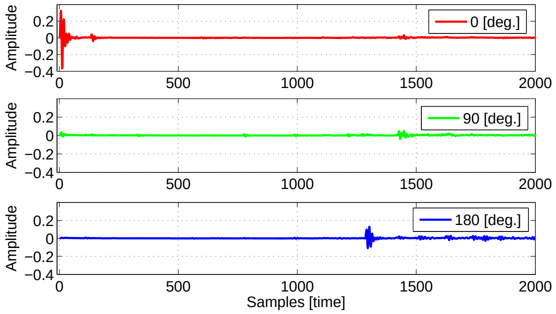

Figure 13. Impulse responses.

TABLE II
OBJECTIVE PARAMETERS.

	Sound source orientation [deg.]							
	0	45	90	135	180	225	270	315
D_{50} [%]	69.5	46.3	41.9	38	42.5	38.2	42.3	45.9
C_{80} [dB]	5.1	1.5	0.8	0.6	1	0.5	1	1.6

sound energy to total sound energy. It is defined in Eq. (8) and expressed as a percentage, where p is the sound pressure. Parameter C_{80} (Clarity) is the ratio of early sound energy to late arriving sound energy. It is defined in Eq. (9) and expressed in dB. This table especially shows the objective differences between the impulse responses from the sound source when it was oriented toward the microphone and oriented elsewhere.

$$D_{50} = \frac{\int_0^{50ms} p^2(t)dt}{\int_0^{\infty} p^2(t)dt} \tag{8}$$

$$C_{80} = 10 \log \left(\frac{D_{80}}{1 - D_{80}} \right) \tag{9}$$

B. Subjective evaluation

We performed a subjective experiment for these impulse responses. For this experiment, we convoluted a speech signal with each impulse response. Therefore, we used these synthesized signals as sound source signals that included orientation. The speech signal was “You’ll mark the book please.” spoken by a male speaker. In this experiment, a subject listened to two continuous speech signals through headphones in random order. The subject reported whether the signal orientations were the same or not. This trial was repeated for all combinations. Each subject participated in all trials four times. The number of subjects was 3.

The accuracy rate for this subjective experiment for each sound source orientation is shown in Fig. 14. In the case where the sound source orientation was 0, 45, 180, 315°, the accuracy rate was approximately 80%. This experimental result shows that a human distinguishes differences in sound source orientation relatively.

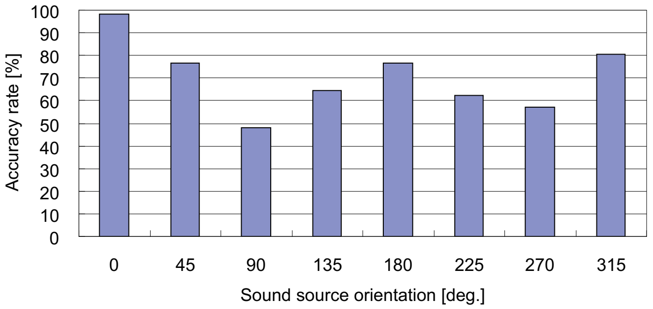

Figure 14. Accuracy rate for subjective evaluation.

VI. Conclusion

In this paper, we proposed a 3D sound field recording and reproducing method that includes sound source orientation. For the recording part, we used a sound source orientation estimation method based on orientation-extended beamforming. We performed experiments to evaluate the orientation estimation method using transfer functions obtained by both practical measurement and calculation using acoustic simulation based on wave theory. The experimental results show that our estimation method has sufficient capability using both measured and calculated transfer functions. For the reproducing part, we evaluated impulse responses versus sound source orientation. The evaluation results revealed significant differences in both objective and subjective scores. This proves that our system achieves higher performance than conventional ones that do not utilize orientation features. Our future work is to extend the sound source orientations to directivities and apply 3D sound field recording and reproducing including sound source orientation estimation to a system that can perform the processing in real time.

References

[1] J. Ohya *et al.*, “Virtual Space Teleconferencing: Real-Time Reproduction of 3D Human Images,” *J. Visual Communication and Image Representation*, vol. 6, no. 1, pp. 1–25, 1995.

[2] T. Betlehem and T. D. Abhayapala, “Theory and design of sound field reproduction in reverberant rooms,” *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2100–2111, 2005.

[3] T. Kimura *et al.*, “Development of Real System in Near 3D Sound Field Reproduction System Using Directional Loudspeakers and Wave Field Synthesis,” *WESPAC*, vol. 164, pp. 1–6, 2009.

[4] F. Asano *et al.*, “Real-time Sound Source Localization and Separation System and its Application to Automatic Speech Recognition,” *EUROSPEECH 2001*, 2001, pp. 1013–1016.

[5] V. Rogijen *et al.*, “Acoustic source number estimation using support vector machine and its application to source localization/separation system,” *IEICE EA*, vol. 102, no. 249, pp. 25–30, 2002.

[6] T. Okamoto *et al.*, “Toward an editable sound-space system using high-resolution sound properties,” *Proc. IWPASH*, P29, 2009.

[7] H. Nakajima *et al.*, “Real-time sound source orientation estimation using a 96 channel microphone array,” *IROS-2009*, 2009, pp. 676–683.

[8] H. Nakajima *et al.*, “An easily-configurable robot audition system using histogram-based recursive level estimation,” *IROS-2010* (to be appeared).

[9] H. Tsuru and R. Iwatsu, “Accurate numerical prediction of acoustic wave propagation,” *Int. J. Adapt. Control Signal Process.*, vol. 24, pp. 128–141, 2010.

[10] S. K. Lele, “Compact finite difference scheme with spectral-like resolution,” *J. Comput. Phys*, vol. 103, pp. 16–42, 1992.