Interdisciplinary Information Sciences Vol. 18, No. 2 (2012) 71–82
©Graduate School of Information Sciences, Tohoku University
ISSN 1340-9050 print/1347-6157 online
DOI 10.4036/iis.2012.71

3D Spatial Sound Systems Compatible with Human's Active Listening to Realize Rich High-Level *kansei* Information

Y. SUZUKI[1,2,*], T. OKAMOTO[3], J. TREVINO[1,2], Z.-L. CUI[1], Y. IWAYA[4], S. SAKAMOTO[1,2], and M. OTANI[5]

[1]*Research Institute of Electrical Communication, Tohoku University, Sendai 980-8577, Japan*
[2]*Graduate School of Information Sciences, Tohoku University, Sendai 980-8577, Japan*
[3]*National Institute of Information and Communications Technology, Kyoto 619-0288, Japan*
[4]*Faculty of Engineering, Tohoku Gakuin University, Tagajo, Miyagi 985-8537, Japan*
[5]*Faculty of Engineering, Shinshu University, Nagano 380-8553, Japan*

Received June 5, 2012; final version accepted August 27, 2012

In future communications with rich high-level *kansei* information, such as presence, verisimilitude, realism and naturalness, the role of sound is extremely important to enhance the quality and versatility of communications because sound itself can provide rich semantic and emotional information. Moreover, sound (auditory information) has good synergy effects with pictures (visual information). In this paper, we introduce our recent research results toward capturing and synthesizing comprehensive 3D sound space information as well as a high-definition 3D audio-visual display realizing strict audio-visual synchronization. We believe that these systems are useful to advance universal communications, which require particularly high-quality and versatile communications technologies for all.

KEYWORDS: communications, *kansei*, spatial audio, active listening, auditory display

1. Introduction

In these decades, the communications technologies have been tremendously advancing, particularly from a quantitative viewpoint. One of the most important reasons for communications must be the realization of high-quality life standards through interactive information exchanges. In this context, communications technologies should aim at qualitative advancements. For us, one key phrase pointing towards this goal is "realization of communications with high-level *kansei* information." *Kansei*, originally a Japanese word, has been accepted as an English word to express the comprehensive set of emotions that arise in evaluating circumstances and things. Typical examples of high-level *kansei* information are presence, verisimilitude, realism, and naturalness. If we could realize communication systems that convey such high-level *kansei* information richly and properly, they would enable us to experience distant places as if we were there, or feel as if distant objects were nearby. The technologies enabling such a feeling would be useful to enhance the quality and versatility of communications to a considerable degree [1, 2]. In such advanced and natural communications, the role of sound is extremely important because sound itself can provide rich semantic and emotional information, i.e. *kansei* information, and sound (auditory information) has good synergy effects with pictures (visual information). It is therefore important to capture, transmit, and synthesize comprehensive three-dimensional (3D) sound space information of a remote place to a local site in such communications.

With this motivation, we have been keenly developing high-definition 3D spatial audio systems as reviewed in later sections of this paper [3–7]. In this series of our efforts, we set our goal to realize a system we call an editable spatial sound field system [8]. In an editable spatial sound field system, various attributes of a specific sound field such as sound source position, sound generated from the sound source, sound source directivity, and reverberations are decomposed precisely. Then, the attributes can be edited, that is, their properties can be specified (or selected) arbitrarily, so that the output sounds are recomposed using the specified/selected attributes. Ideally speaking, any desired sound in any acoustic environment might be produced and listened to using this system. Section 3 of this paper presents a more detailed introduction of the notion of this editable spatial sound field system.

In developing such systems, we are acutely conscious of a fact that spatial hearing is a multimodal perception consisting of not only hearing but also self-motion perception. Furthermore, we are putting equal stress on systems to render spatial sound information (auditory displays) and systems to capture 3D spatial sound information (spatial sound

* Corresponding author. E-mail: yoh@riec.tohoku.ac.jp

acquisition systems). Fewer studies have examined technologies for 3D sound acquisition than those for auditory displays. However, in our view, technologies of these two kinds are the right and left wheels of the tractor of high-definition 3D spatial audio technology.

In §2 and later, we review our recent research results conducted with these ideas. In §2, we first introduce our attempts to investigate human three-dimensional (3D) spatial perception using a high-definition 3D binaural auditory display. In §3, systems based on a 157 surrounding microphone array are introduced for realizing editable sound field system. In §4, our high-definition 3D spatial sound acquisition system, called SENZI (Symmetrical object with ENchased ZIllion microphones), is introduced. This system is based on a microphone array on a human-head-sized solid sphere with numerous microphones on its surface. It can comprehensively record and/or transmit accurate sound-space information to a distant place. In §5, high-definition 3D spatial sound acquisition systems and auditory displays based on higher-order Ambisonics (HOA) are introduced along with their basic underlying theory. The order of this higher-order Ambisonic auditory display is five, which is the highest order, and therefore the highest precision realized to date. Moreover, a 3D audio-visual display consisting of this fifth-order Ambisonic auditory display and a 3D projection display is introduced. A few concluding remarks are presented in §6.

2. Effect of Head Movement in Three-Dimensional Spatial Hearing

To realize future advanced communications as mentioned in the previous section, it is important to recall that we humans are active creatures, moving through the environment to acquire accurate spatial information. For instance, in terms of spatial hearing, humans usually make slight head and body movements unconsciously, even when trying to keep still while listening. Actually, such movement is known to be effective for improving the precision of auditory spatial recognition [9–14]. We designate this style of listening as active listening [15]. Therefore, it is particularly important that 3D spatial sound systems be compatible with a listener's movement, at least to a listener's head rotation. Three dimensional auditory displays and spatial sound acquisition systems matching the motions of active listening are therefore eagerly sought for use in future communications.

2.1 Perception of ambient sound

In the 3D auditory display system, to render a sound associated with a specific object (target sound), such as voice, sounds of musical instruments located at a specific position, the sound is convolved with head-related transfer functions (HRTFs) corresponding to the position information of the sound [16]. An HRTF represents sound transmission characteristics from a sound source to a listener's ear.

In our daily life, we not only hear one or a few specific sound sources: we are surrounded by sounds arriving at our ears after being emitted from the sound source and interacting with the environment through various physical phenomena such as reflection, reverberation, diffraction, and Doppler shift. However, 3D auditory displays, especially those which require real-time processing, usually cannot render these phenomena. Consequently, a listener sometimes feels unnatural with virtual sound space presented by a 3D auditory display system. Accordingly, it seems necessary to add ambient sound to rendering of the auditory display.

Therefore we investigated the effects of ambient sounds using subjective evaluations in a virtual environment and propose a rendering method for ambient sound [17]. First, after examining frequency characteristics of several actual environmental noises, red noise, noise with a frequency spectrum of $-6\,\mathrm{dB/oct}$, was selected as an optimal noise for synthesizing ambient sounds. Next, multiple red noise sources with uncorrelated phase characteristics as well as a point sound source as a target sound were spatially distributed by convolution with head-related impulse responses (HRIRs), which are the time-domain representation of HRTFs.

In the first experiment, effects of spatialized ambient sound on the reality of sound space perception with head movement was examined. Here, only target sound was responsive to head movement. The results of the experiment are presented in Fig. 1. This figure shows that presentation of the spatialized ambient noise significantly improved a perception of realism (Fig. 1). In real world conditions, ambient sound usually exhibit specific spatial patterns. Therefore, in the next experiment, whether presentation of ambient noise with some spatial pattern could enhance realism of the presented sound space [17]. In this experiment, ambient noise sources have variations in their sound pressure levels depending on their position to have a non-uniform spatial pattern. In this method, both target and ambient noise were responsive to head movement. That is, not only the position of the target sound but also the spatial pattern of the ambient noise sources, in terms of the absolute coordinate, was kept unchanged against the listener's head movement. Experimental results show that the perceived realism when listeners move their head is statistically significantly higher than the perceived realism when they did not move their head [$t(8) = 2.48$, $p < 0.05$]. This indicates that perceived realism of virtual sound space rendered along with artificial ambient sounds with a non-uniform spatial patterns is significantly improved if both a target sound and ambient sounds are responsive to a listener's head movement.

2.2 Effect on sound localization in the median plane

As described above, head movement during listening in a sound space can enhance the realism, and therefore the

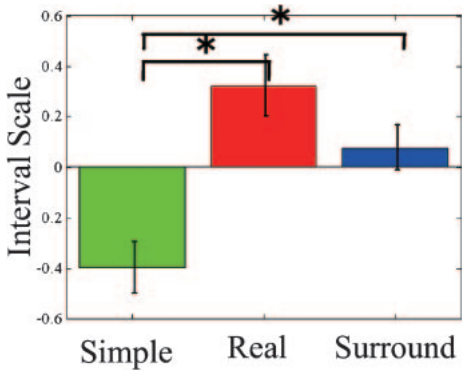


Fig. 1. Mean interval scales of the perceived realism. Participants perceived greater realism for in surround condition than in simple condition although no significant difference was found between real and surround conditions.

high-level *kansei* information such as sense-of-presence and verisimilitude, of the perceived sound space. For example, Kawaura *et al.* investigated sound localization accuracy using a 3D auditory display based on HRTF synthesis dynamically reflecting listener's horizontal head rotation sensed by a potentiometer attached at the top of the listener's head [11, 12]. The accuracy of the sound localization, particularly the distance perception and front-back confusion are significantly improved by dynamically change synthesized HRTF to reflect the head rotation. Later, Toshima *et al.* investigated sound localization accuracy using TeleHead in a horizontal plane and median plane [18]. TeleHead is an avatar robot of a human listener who would be in a different place from the TeleHead and listening to sound space via microphones installed at the ear positions of TeleHead. They showed that horizontal head rotation plays the most important role in localizing sound in the horizontal plane [9]. Therefore, to examine the effect of horizontal head rotation specifically in greater detail, we developed a simplified TeleHead whose head moves synchronously following the listener's horizontal head rotation.

Using implemented TeleHead, we investigated the sound localization accuracy in the median plane because localization of elevation angles is very important to render a three-dimensional sound space. Figure 2 shows the experimental setup.

Sound signals at the TeleHead robot's two ears in an anechoic room are captured and reproduced for a listener in a remote soundproof room. The robot rotation was controlled so that the rotation ratio between the listener and the robot varied from 0.05 to 1.0.

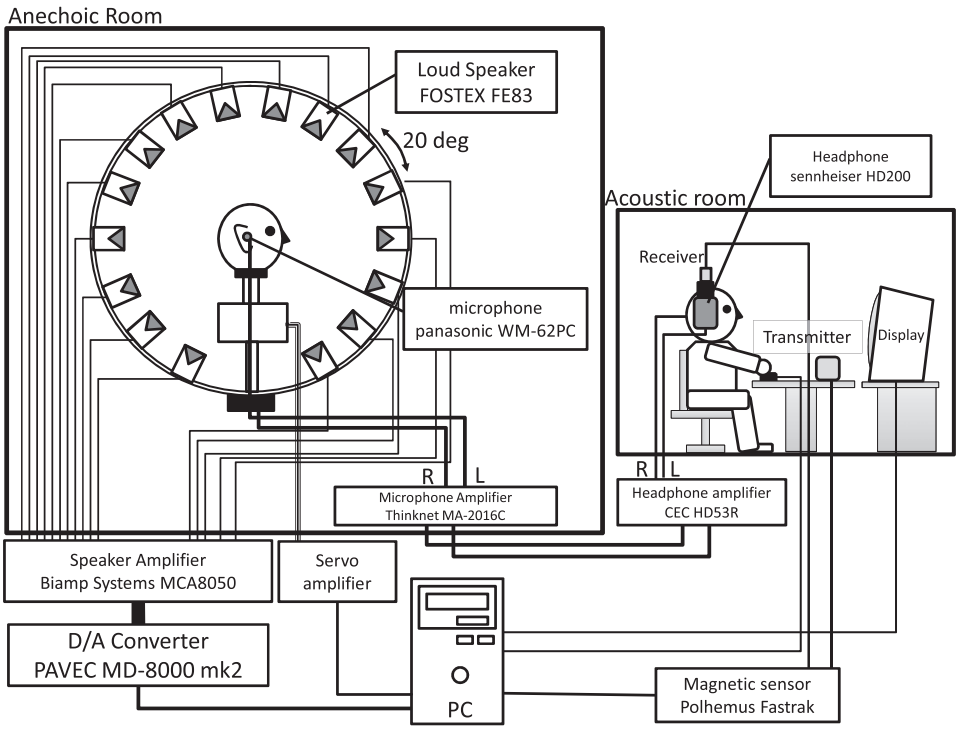
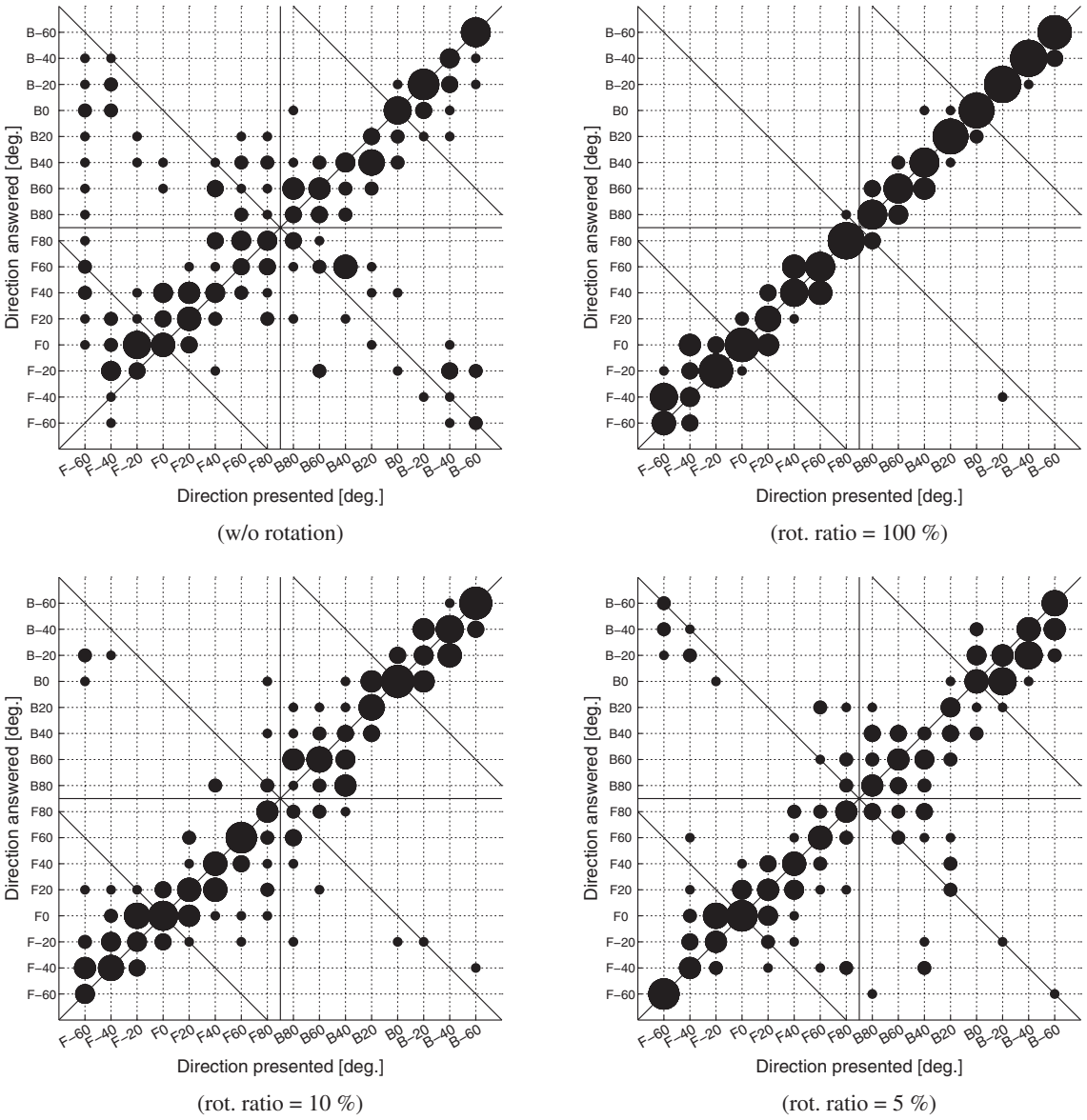


Fig. 2. Experimental setup.



(w/o rotation) (rot. ratio = 100 %)

(rot. ratio = 10 %) (rot. ratio = 5 %)

Fig. 3. Results of experiment (Listener 1).

Results of the experiment of Listener 1 are depicted in Fig. 3 as examples. These results show that frequent front-back confusions are observed when an avatar robot does not respond to the listener's head rotation. In contrast, when the robot rotates in-phase to the listener's head rotation to provide dynamic sound localization cues, front-back confusions are significantly suppressed, irrespective of the rotation magnitude. Consequently, the sound localization accuracy can be improved considerably if the robot rotates in a remote site, even if the robot head rotation magnitude is as little as 5% of the listener's head rotation. It should be also mentioned that listeners could not notice that the sound localization was dynamically controlled to reflect the head rotation when the magnitude of the robot head rotation was less than or equal to 10% of the listener's head rotation. This suggests that head movement is implicitly utilized to stabilize sound localization and that not the magnitude itself but just the direction of the head rotation is perceptually important to improve sound localization in the median plane.

3. Efforts Toward Editable Spatial Sound Field Systems

In the future of acoustic signal processing, numerous microphones can be used inexpensively and a whole sound field in a room can thereby be recorded accurately using a microphone array consisting of the microphones. As a consequence of such signal processing, the sound field can be decomposed further into some attributes such as original sound source signals, sound source positions, directivity of sound sources, reflections, and reverberation. If such possibilities are realized, then editing of the sound field would become highly versatile. Not only the original sound field but also a modified sound field can be synthesized by modifying and exchanging these attributes. We designate such a system as an editable sound field system. It is important to conduct necessary studies to realize such a system

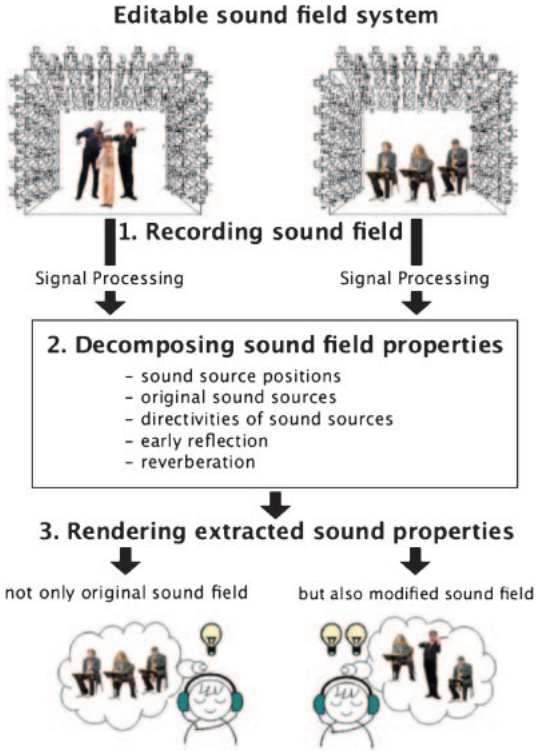


Fig. 4. Editable spatial sound field systems.

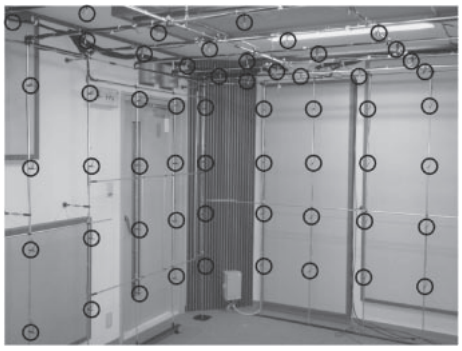


Fig. 5. Appearance of the surrounding microphone array (each black circle represents a microphone).

because previous efforts on sound field reproduction [19] as well as sound field acquisition are insufficient to realize modified sound fields such as those described above (Fig. 4). To record sound information necessary to realize an editable sound field system, we constructed a test-bed room for sound acquisition in which a microphone array consisting of 157 microphones (Type 4951; Bruel and Kjær) is installed on all four walls and the ceiling of the room. We designate this as a surrounding microphone array. All microphones are installed 30 cm inside from all four walls and the ceiling using pipes. They are separated from one another by 50 cm. The surrounding microphone array is shown in Fig. 5. We introduced a recording system for this microphone array to enable synchronous recording of 157 channels at the sampling frequency of 48 kHz with the linear PCM audio format [3]. To date, we have developed estimation methods of sound source positions [3], directivity of a sound source [20], and a dereverberation method [21] using this array.

4. Spherical Microphone Array to Capture 4π Spatial Sound Information (SENZI)

4.1 Basic theory

We have proposed an acquisition system of 3D sound-space information that can record and/or transmit accurate sound-space information to a distant place using a microphone array on a human-head-sized solid sphere with numerous microphones on its surface. We designate this system as SENZI, which is an acronym for Symmetrical object with ENchased ZIllion microphones [4]; here SENZI also means one thousand ears literally in Japanese. The system

can sense 3D sound-space information comprehensively. That is, correct spatial information from all directions is available for any listener orientation.

In the system, the signal from each microphone is simply weighted and summed. Here, the weight is changed according to a listener's 3D head movement, which is an important cue to perceive a 3D sound space with high sense-of-presence [14]. Therefore, accurate 3D sound-space information is acquired irrespective of the listener orientation.

The optimal set of each weight in the frequency domain $\mathbf{z}_f$ for each microphone is calculated based on the least-squares algorithm.

$$\mathbf{z}_f = \mathbf{H}^{+}_{\mathrm{SENZI},f}\mathbf{H}_{\mathrm{listener},f}. \tag{1}$$

Therein, $\mathbf{H}_{\mathrm{SENZI},f}$ and $\mathbf{H}_{\mathrm{listener},f}$ respectively denote the transfer function between a microphone on SENZI and a direction and the listener's head-related transfer function (HRTF [16]) from a direction.

In that equation, $\mathbf{H}^{+}$ stands for the pseudo-inverse of matrix $\mathbf{H}$. When $\mathbf{z}_f$ is calculated, it is necessary to select directions that are incorporated into the calculation. The selected directions are designated as controlled directions. As in a real environment, sound waves come from all directions, including directions that are not incorporated into calculations. We will refer to these as uncontrolled directions. To capture accurate sound information, transfer functions not only for the controlled directions but also for all directions including the uncontrolled directions should be synthesized appropriately. To realize this, the number of microphones, the arrangement of the microphones on the object, and the shape of the object must be optimized [4].

SENZI has one advantageous characteristic over other 3D spatial sound information acquisition system such as ordinary dummy head and TeleHead. That is, with SENZI, once the sound signals from all the microphones are recorded, appropriate 3D spatial sound can be reproduced for any head orientation. This means that precise 3D spatial hearing perception can be realized at any time at any place, at or after the recording with SENZI.

4.2 Implementation of 252 channel real-time SENZI

We designed and produced an actual system (Fig. 6). The object radius is 17 cm, which is determined according to the head size. The spherical object has 252 microphones on it: the position of each was calculated based on a regular icosahedron. Each surface of a regular icosahedron was divided into 25 small equilateral triangles to install 252 microphones. A small omnidirectional digital ECM microphone (KUS5147; Hosiden Corp.) is set at the calculated position. The rigid sphere surface was made from epoxy resin using stereolithography. In each digital ECM, the recording signal was sampled at 3.072 MHz and transformed 16 bit data by fourth $\delta\sigma$ modulation. The signal-to-noise ratio (SNR) of the digital ECM is 58 dB (typical value). All 252 digital signals were assembled at the control substrate in the rigid sphere; the dataset was transferred to the control board and the two FPGA boards through two data transfer cables. In the FPGA boards, the dataset was converted to 252 linear PCM signals; the sampling frequency was 48 kHz and 16 bit. We are now enhancing this SENZI system into a real-time one to develop a sound space recording and synthesizing system with it.

5. Acquisition Systems and Auditory Displays based on Higher-order Ambisonics

5.1 Basic theory of Ambisonics

Three dimensional auditory displays that reproduce/render 3D sound space in a specific region have been developed based on the Kirchhof–Helmholtz equation. Wave Field Synthesis (WFS) [24] and Boundary Surface Control (BoSC) [19] are the well known architectures. Alternative approaches, known as Ambisonics [22, 23], can overcome the main shortcomings of Wave Field Synthesis (WFS) and Boundary Surface Control (BoSC). Like WFS, Ambisonics assumes

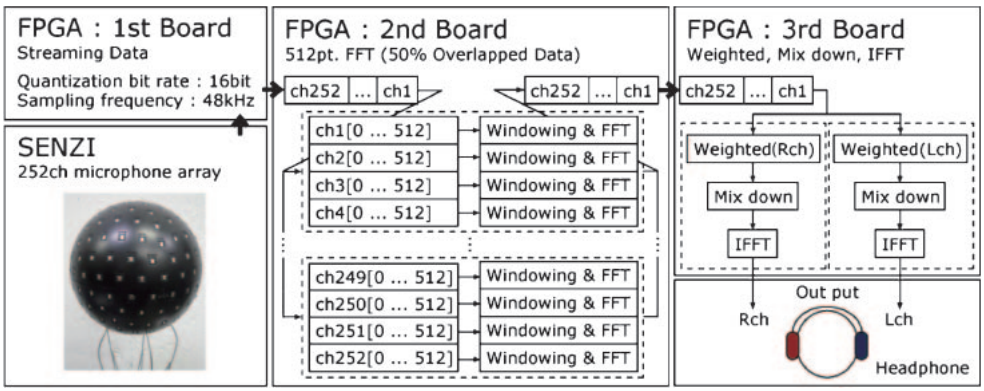


Fig. 6. System scheme of SENZI.

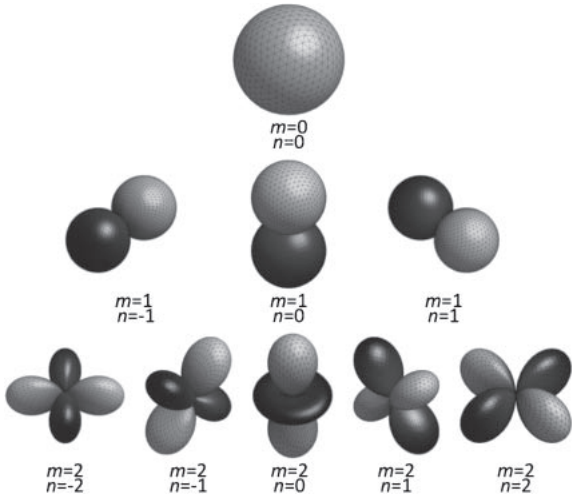


Fig. 7. 3D view of spherical harmonics up to order 2 with usual designation of associated HOA components.

Dirichlet boundary conditions: consequently, it needs no complex hardware such as the extended microphone arrays required by BoSC. However, it divides the space using a closed boundary surface. This allows Ambisonics to present sound from all directions, thereby overcoming the main limitation of WFS. Reproduction of sound fields using Ambisonics requires a surrounding loudspeaker array, with loudspeakers re-creating the sound pressure on the boundary.

By positioning the listener at the center of a surrounding loudspeaker array, Ambisonics can be best described using the spherical coordinate system. Each loudspeaker is identified by its azimuth and elevation angles as seen from the listening point, as well as its distance to the listener. The use of spherical coordinates makes it straightforward to expand the Kirchhoff–Helmholtz integral equation in terms of special functions; this is known as the multipole expansion [25]. Particularly the basis functions used in the multipole expansion are known as the spherical harmonics and are calculable as

$$\mathrm{Y}_m^n(\theta, \varphi) = N_m^n e^{in\theta} P_m^n(\cos\varphi), \tag{2}$$

where θ denotes the azimuth angle, φ is the elevation angle, and N_m^n represents a normalization constant. Figure 7 shows the spherical harmonics from 0-th to second order.

Basic Ambisonic [22] systems use only the first four terms of the multipole expansion, namely the 0-th and the first harmonics shown in Fig. 7. These terms are sufficient to characterize the sound pressure and its spatial derivative fully at one point in space-the center of the array. Most basic Ambisonics, sometimes referred to as first-order Ambisonics, is sufficient for simple applications and has low hardware requirements. Reproducing the sound pressure and its derivative at one point is, however, insufficient for more advanced applications. Additional terms of the multipole expansion must be considered in some tasks, such as accurate representation of sound to human listeners, reproducing extended sound sources, re-creating complex radiation patterns, and supporting multiple listeners within a wide listening region. The use of multipole expansion terms above the first spherical harmonic degree such as the second harmonics shown in Fig. 7 or even higher is Higher Order Ambisonics (HOA) [22, 23], which is also sometimes called Generalized Ambisonics. Still, computational limitations require that the multipole expansion be truncated after a few terms. The particular degree at which the expansion is truncated is known as the Ambisonic order. It is usually determined by consideration of the desired directional resolution and the plausible system size.

By limiting the description of the sound field to only a few expansion coefficients, the amount of information that recording, processing, and reproduction systems need to handle is drastically reduced. Mathematically, the Ambisonic encoding of the sound field $\vec{p}$ can be written as

$$B_{mn}(k) = \int_0^{2\pi} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{p(\vec{r}, k)}{j_m(kr)} \mathrm{Y}_{mn}(\theta, \varphi) \sin\varphi \, d\varphi \, d\theta, \tag{3}$$

where k denotes the wavenumber, and where j_m are spherical Bessel functions that characterize the radial component of the sound field in spherical coordinates.

The encoded sound field of eq. (3) does not depend explicitly on the distribution of loudspeakers to be used in its reproduction. This is an important advantage of Ambisonics; it can be reproduced using any surrounding loudspeaker array. However, a decoding stage is necessary to derive loudspeaker signals from the set of multipole expansion coefficients. If the loudspeaker array is regular, that is, if its loudspeakers sample the sphere at uniform intervals, then decoding can be done by taking the pseudo-inverse [26] of a matrix whose elements are the spherical harmonic functions evaluated at positions of the loudspeakers in the target array. Unfortunately, regular loudspeaker distributions

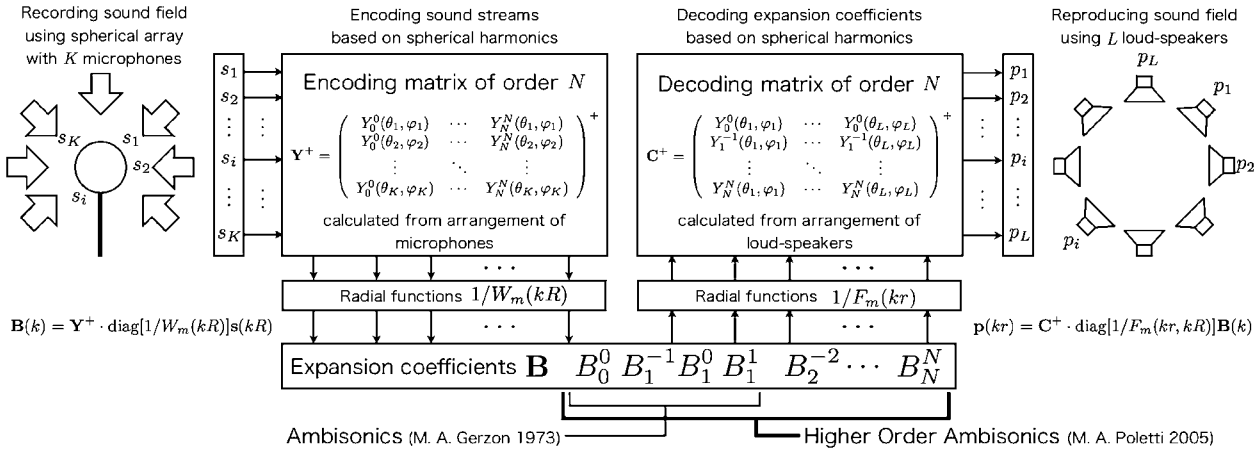


Fig. 8. Sound field recording and reproduction schema based on Higher-order Ambisonics.

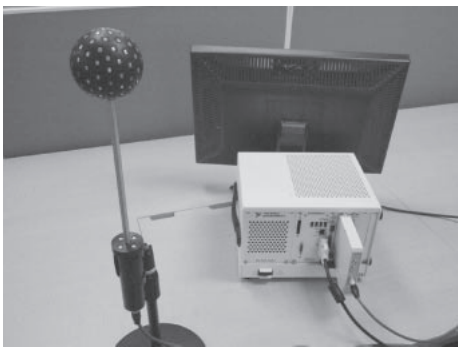


Fig. 9. External look of the 121-ch Ambisonic microphone array system.

are impractical for most applications and the pseudo-inverse can yield numerically unstable results if an irregular loudspeaker distribution is used [27].

5.2 121 ch Ambisonics microphone array

For reproducing an actual sound field based on HOA, development of high-definition recording systems is important, too. Usually, an HOA recording system is implemented as a spherical microphone array. For example, 64 channels of microphones were used in a previous implementation of spherical microphone arrays [28]. We implemented a spherical HOA recording array using 121 digital omnidirectional electric condenser microphones (ECMs) for recording actual sound fields more accurately. Figures 8 and 9, respectively, show a block diagram of the implemented recording system and the appearance of the actual recording system. The basic hardware system architecture and units, including the microphones, are the same as the 252-ch SENZI system. The recording signal was sampled at 3.072 MHz and transformed 16 bit data by fourth $\Delta\Sigma$ modulation. All 121 digital signals were assembled at the control substrate in the rigid sphere; the dataset was transferred to the control board and the FPGA board through a data transfer cable. In the FPGA board, the dataset was converted to 121 16-bit, linear PCM signals at the sampling frequency of 48 kHz and 16 bit. In this system, all 121 sound signals can be recorded completely and synchronously [5].

5.3 Enhancement of Ambisonic display to broaden its sweet spot

We implemented a HOA display using an irregular, surrounding, 157-channel loudspeaker array, shown in Fig. 10, to present spatial audio [29]. The particular loudspeaker distribution of this array does not describe a uniform sampling of all angles. It is therefore unfit for HOA reproduction if a standard decoder is used. We have proposed a new Ambisonic decoder that can weaken the requirement of a regular array [6], and applied it for the reproduction of sound fields using the 157-channel array. Our Ambisonic decoding proposal is outlined in Fig. 11. It is divisible into two stages: a standard decoder for lower order channels and a radial stabilization stage. The first stage makes use of the pseudo-inverse, as previously outlined in §5.1. However, it does not consider all terms in the Ambisonic encoding. Rather, it ensures numerical stability by decoding only those terms which result in a well-conditioned matrix of spherical harmonics. The second stage improves the presentation of sound by attempting to enlarge the listening region. For this, our proposal is to look for the decoding gains that will minimize the magnitude of the radial derivative of the reconstruction error: we impose the constraint of a smooth error field. The decoding gains are calculable as

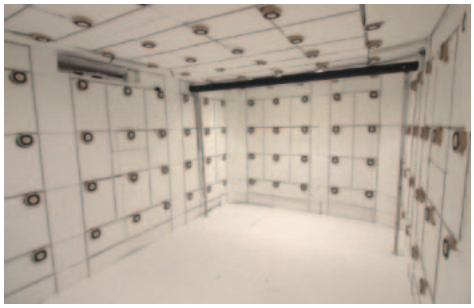


Fig. 10. Appearance of the surrounding loudspeaker array.

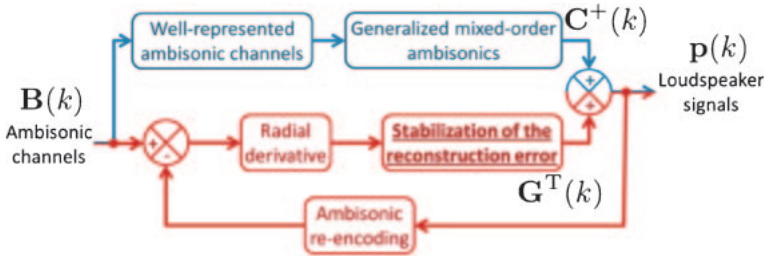
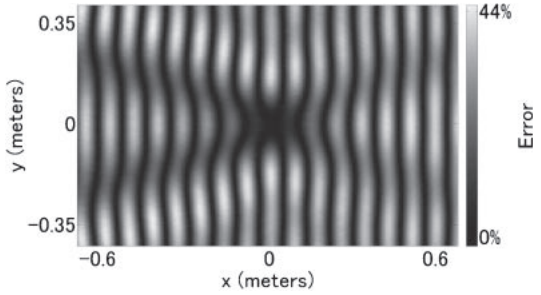
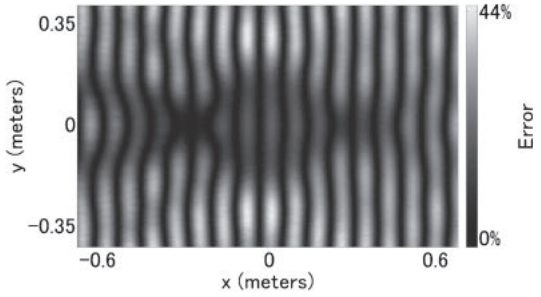


Fig. 11. Proposed Ambisonic decoder.



(a) Reconstruction error when using the pseudo-inverse method.



(b) Reconstruction error when using the proposed method.

Fig. 12. Reconstruction error for a 2 kHz plane wave incident from the right.

$$\mathbf{G} = \arg\min_{\mathbf{G}} \left| \frac{\partial}{\partial r} \left[\tilde{\psi}_k(r,\theta,\varphi) - \phi_k(r,\theta,\varphi) - \sum_s \sum_{m=0}^{N} \sum_{n=-m}^{m} G^s_{mn}(k) \frac{e^{-ik|\vec{r}-\vec{r}_s|}}{|\vec{r}-\vec{r}_s|} \mathrm{Y}_{mn}(\theta_s,\varphi_s) \right] \right|, \tag{4}$$

where ψ denotes the target sound field, ϕ the sound field re-created in the first stage and s is the loudspeaker index.

To evaluate our proposal, we conducted a computer simulation of the 157-channel loudspeaker array. We present the results for the reconstruction error of a 2 kHz plane wave in Fig. 12. As an estimate of the reconstruction error, we calculated the RMS value of the error fields for both decoders within a 17-cm diameter sphere. The standard decoder results in an error of 1.63 dB, whereas our proposal achieves a lower error of 0.95 dB in the region of interest.

Furthermore, our simulation results show that, although the accuracy at the precise center of the array is lower when using our proposal, the size of the listening region, i.e., the area in which the error remains reasonably low, is larger.

5.4 Implementation of a high-definition 3D audio-visual display

Based on the a 157-loudspeaker array, we actually implemented an HOA auditory display [30]. The decoding order in our system was five, the greatest in the world at present, which is especially useful for localization experiments and listening tests because the reproduction of HOA assumes free field [30].

It is vital in a multichannel audio system to achieve precise synchronicity because human hearing system can distinguish the arrival time difference between the two ears as short as 10 μs as the difference of the perceived sound image localization. Thus, to realize complete synchronization of all 157 audio signals, we introduced the MADI system, controlled by a single computer, to our system. A MADI system can control up to 64 input–output audio channels in complete synchrony at the single-sample level using only a single connection. The MADI PCI express interface (HDSPe MADI; RME) can receive up to three MADI connections. Therefore, up to 192 audio signals can be controlled by a single PC using one MADI interface card. For controlling 157 audio signals using only one PC, we have introduced audio control software (Pd-0.42-5; Pure Data) and five D/A converters (M32-DA; RME) that have 32 output channels in one unit. Clock synchronization is achieved using a master clock generator (Nanosync HD; Rosendahl Studiotechnik GmbH) connected to each MADI connection. After installing the driver of the MADI interface in the control PC, the Mac PC's audio driver (Core Audio) was set for integrating of three MADI systems into one wrapped audio device that has 192 outputs. To reduce the latencies, another audio driver, Jack Pilot was adopted. Jack Pilot is commonly used to integrate input–output devices into one device from an application. Another advantage of Jack Pilot is that the audio input–output processes can be distributed among the available CPU cores. Consequently, the controlling application (Pure Data) and the audio output process are divisible. The Jack Pilot buffer size was set to 512 kbyte and the buffer size of PD was set as 50 ms. Using this system, a fifth-order HOA was constructed. A 3D audio-visual display was implemented using the sound subsystem described in the preceding paragraphs. It is combined with a 3D projection display based on stereovision. Figure 13 shows a block diagram of the developed system.

The system latency difference between the drawing on the screen and 157-audio stream was 51 samples (= 1.1 ms). This value is much less than the detection threshold of audio-visual asynchrony by human observers [31]. Therefore, various 3D multimodal contents consisting of a sound field recorded by a HOA microphone [5] or a simulated sound field [32] combined with a 3D drawing captured using a stereo camera or a simulated one can be reproduced using this system with rich high-level *kansei* information.

6. Concluding Remarks

This paper reviews our recent efforts at realizing high-definition 3D spatial audio systems consisting of 3D auditory displays and sound information acquisition systems. In this series of efforts, we have aimed at achieving an editable

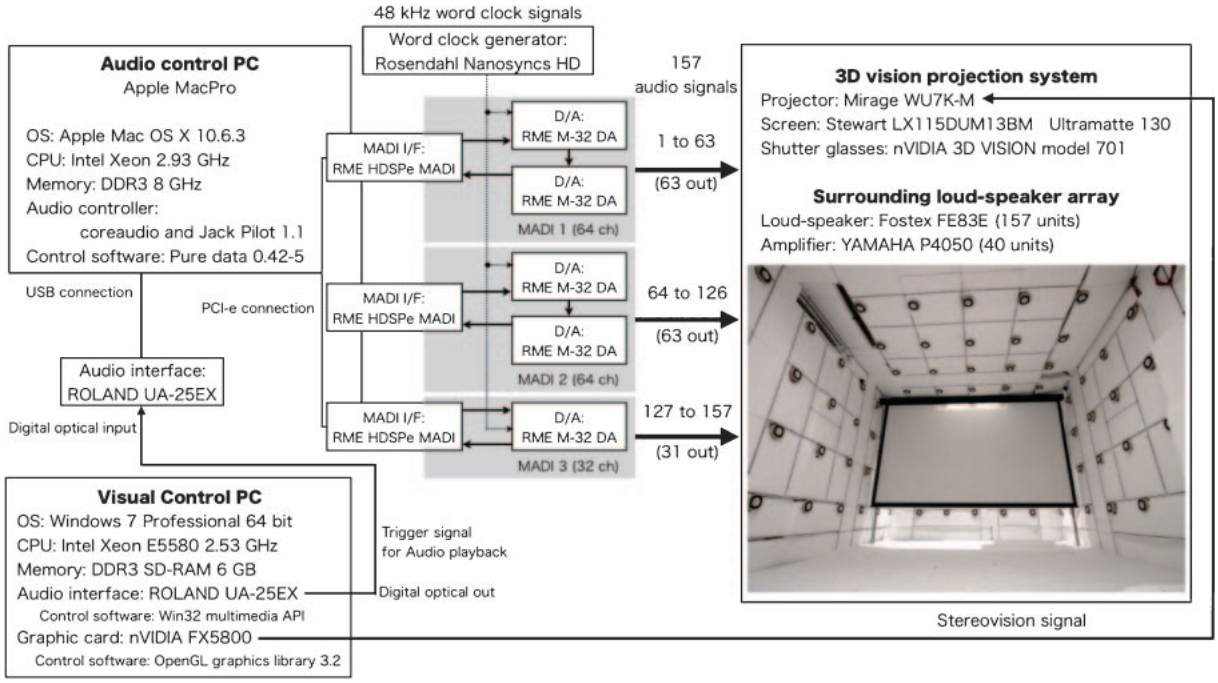


Fig. 13. 3D audio-visual reproduction system.

sound field system with a notion of "active listening." An editable sound field system as well as the notion of active listening are important to create and/or reproduce any sound in any acoustic environment used for listening. We have been developing high-definition 3D spatial sound acquisition systems with more than 100 channels and high-definition 3D auditory displays of several kinds. Good 3D sound acquisition systems and good 3D auditory displays must be the two wheels of the powerful tractor of high-definition 3D spatial audio technology. In the last part of this paper, we introduced a 3D audio-visual display applying a 3D projection display and a higher-order Ambisonic display with the highest order, and therefore the highest precision realized in the world today.

As another series of study of ours, we have been putting our efforts to clarify how high-level *kansei* information such as presence and verisimilitude are evoked [1, 2]. The results show that the sense of presence is multidimensional perception which strongly depends on the total intensity of the given information and increases as the total intensity increases. On the other hand, the sense of verisimilitude is mainly governed by foreground information and peaks out at a suitable intensity. In future, therefore, to realize yet advanced communications with rich high-level *kansei* information, we should be aware of such characteristic difference among the aspects of high-level *kansei* information as we found between presence and verisimilitude. We believe that our recent research results reviewed in this paper must be useful as the bases for developing high-definition multimedia systems that carry rich high-level *kansei* information to advance future communications towards this direction.

Acknowledgment

Parts of the research were supported by Tohoku University GCOE program CERIES, Grants-in-Aid for Specially Promoted Research (No. 19001004) to SY from MEXT, and for SCOPE (No. 082102005) to SS from MIC, Japan. The authors wish to thank the colleagues who participated in the grant programs for their intensive discussions. In particular, we would like to single out Jiro Gyoba as a leading contributor during our studies of high-level *kansei* information.

REFERENCES

[1] Teramoto, W., Yoshida, Y., Asai, N., Hidaka, S., Gyoba, J., and Suzuki, Y., "What is "Sense of Presence"? A non-researcher's understanding of sense of presence," *Trans. Virtual Real. Soc. Jpn.*, **15**: 7–16 (2010) (in Japanese).

[2] Teramoto, W., Yoshida, Y., Hidaka, S., Asai, N., Gyoba, J., Sakamoto, S., Iwaya, Y., and Suzuki, Y., "Spatio-temporal characteristics responsible for high 'Vraisemblance'," *Trans. Virtual Real. Soc. Jpn.*, **15**: 483–486 (2010) (in Japanese).

[3] Okamoto, T., Nishimura, R., and Iwaya, Y., "Estimation of sound source positions using a surrounding microphone array," *Acoust. Sci. & Tech.*, **28**: 181–189 (2007).

[4] Sakamoto, S., Kodama, J., Hongo, S., Okamoto, T., Iwaya, Y., and Suzuki, Y., "Effects of microphone arrangement on the accuracy of a spherical microphone array (SENZI) in acquiring high-definition 3D sound space information," *Book Chapt. Princ. and Appl. on Spatial Hearing*, 314–323 (2011).

[5] Okamoto, T., Iwaya, Y., Sakamoto, S., and Suzuki, Y., "Implementation of higher order Ambisonics recording array with 121 microphones," *Proc. 3rd Student Organizing Int. Mini-Conf. on Inf. Electr. Syst.*, 71–72 (2010).

[6] Trevino, J., Okamoto, T., Iwaya, Y., and Suzuki, Y., "Higher order Ambisonic decoding method for irregular loudspeaker arrays," *Proc. ICA 2010* (2010).

[7] Okamoto, T., Cui, Z.-L., Iwaya, Y., and Suzuki, Y., "Implementation of a high-definition 3D audio-visual display based on higher-order Ambisonics using a 157-loudspeaker array combined with a 3D projection display," *Proc. IEEE IC-NIDC 2010*, 179–183 (2010).

[8] Okamoto, T., Iwaya, Y., and Suzuki, Y., "Estimation of high-resolution sound properties for realizing an editable sound-space system," *Book Chapt. Princ. and Appl. on Spatial Hearing*, 407–416 (2011).

[9] Thurlow, W. R., Mangels, J. W., and Runge, P. S., "Head movements during sound localization," *J. Acoust. Soc. Am.*, **42**: 489–493 (1967).

[10] Thurlow, W. R., and Runge, P. S., "Effect of induced head movements on localization of direction of sound," *J. Acoust. Soc. Am.*, **42**: 480–488 (1967).

[11] Kawaura, J., Suzuki, Y., Asano, F., and Sone, T., "Sound localization in headphone reproduction by simulating transfer function from the sound source to the external ear," *J. Acoust. Soc. Jpn.*, **45**: 756–766 (1989) (in Japanese).

[12] Kawaura, J., Suzuki, Y., Asano, F., and Sone, T., "Sound localization in headphone reproduction by simulating transfer function from the sound source to the external ear," *J. Acoust. Soc. Jpn. (E)*, **12**: 203–216 (1991).

[13] Perrett, S., and Noble, W., "The effect of head rotations on vertical plane sound localization," *J. Acoust. Soc. Am.*, **102**: 2325–2332 (1997).

[14] Iwaya, Y., Sukuki, Y., and Kimura, D., "Effects of head movement on front-back error in sound localization," *Acoust. Sci. & Tech.*, **24**: 322–324 (2003).

[15] Suzuki, Y., "Auditory Displays and Microphone Arrays for Active Listening," Keynote lecture, 40th Int. AES Conf. (2010).

[16] Blauert, J., *Spatial Hearing*, Cambridge, MIT Press (1983).

[17] Yairi, S., Iwaya, Y., Kobayashi, M., Otani, M., Suzuki, Y., and Chiba, T., "The effect of ambient sounds on the quality of a 3D virtual sound space," *Proc. IIH-MSP 2009*, 1122–1125 (2009).

[18] Iwaki, T., and Aoki, S., "Sound localization during head movement using an acoustical telepresence robot: TeleHead," *Advanced Robotics*, **23**: 289–304 (2009).

[19] Ise, S., "A principle of sound field control based on the Kirchhoff–Helmholtz integral equation and the theory of inverse

systems," *ACUSTICA — Acta Acustica*, **85**: 78–87 (1999).
[20] Okamoto, T., Iwaya, Y., and Suzuki, Y., "Blind directivity estimation of a sound source in a room using a surrounding microphone array," *Proc. ICA 2010* (2010).
[21] Okamoto, T., Iwaya, Y., and Suzuki, Y., "Wide-band dereverberation method based on multichannel linear prediction using prewhitening filter," *Appl. Acoust.*, **73**: 50–55 (2012).
[22] Gerzon, M. A., "Periphony: with-height sound reproduction," *J. Audio Eng. Soc.*, **21**: 2–10 (1973).
[23] Poletti, M. A., "Three-dimensional surround sound systems based on spherical harmonics," *J. Audio Eng. Soc.*, **53**: 1004–1025 (2005).
[24] Berkhout, A. J., deVries, D., and Vogel, P., "Acoustic control by wave field synthesis," *J. Acoust. Soc. Am.*, **93**: 2764–2778 (1993).
[25] Jackson, J. D., *Classical Electrodynamics, Third Edition*, ed. Wiley (1998).
[26] Golub, G. H., and Van Loan, C. F., *Matrix Computations, Third Edition*, ed. Baltimore (1996).
[27] Zotter, F., "Sampling Strategies for Acoustic Holography/Holophony on the Sphere," *Tech. Rep. Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz* (2008).
[28] Zotkin, D. N., Duraiswami, R., and Gumerov, N. A., "Plane-wave decomposition of acoustical scenes via spherical and cylindrical microphone array," *IEEE Trans. Audio, Speech, Lang. Process.*, **18**: 2–16 (2010).
[29] Okamoto, T., FG Katz, B., Noisternig, M., Iwaya, Y., and Suzuki, Y., "Implementation of real-time room auralization using a surrounding 157 loudspeaker array," *Book Chapt. Princ. and Appl. on Spatial Hearing*, 373–382 (2011).
[30] Okamoto, T., Cabrera, D., Noisternig, M., FG Katz, B., Iwaya, Y., and Suzuki, Y., "Improving sound field reproduction in a small room based on higher-order Ambisonics with a 157-loudspeaker array," *Proc. 2nd Int. Symp. on Ambisonics and Spherical Acoust.* (2010).
[31] Spence, C., Baddeley, R., Zampini, M., James, R., and Shore, D. I., "Multisensory temporal order judgments: When two locations are better than one," *Percept. Psychophys.*, **65**: 318–328 (2003).
[32] Noisterning, M., FG Katz, B., Siltanen, S., and Savioja, L., "Framework for real-time auralization in architectural acoustics," *Acta Acustica United with Acustica*, **94**: 1000–1015 (2006).