Acoust. Sci. & Tech. **36**, 6 (2015)

PAPER

Sound-space recording and binaural presentation system based on a 252-channel microphone array

Shuichi Sakamoto[1,*], Satoshi Hongo[2], Takuma Okamoto[3], Yukio Iwaya[4] and Yôiti Suzuki[1]

[1]*Research Institute of Electrical Communication and Graduate School of Information Science, Tohoku University, 2–1–1 Katahira, Aoba-ku, Sendai, 980–8577 Japan*
[2]*Department of Design and Computer Applications, Sendai National College of Technology, 48 Nodayama, Medeshima-Shiote, Natori, 981–1239 Japan*
[3]*National Institute of Information and Communications Technology, 3–5 Hikari-dai, Seika-cho, Soraku-gun, Kyoto, 619–0289 Japan*
[4]*Faculty of Engineering, Tohoku Gakuin University, 1–13–1 Chuo, Tagajo, 985–8537 Japan*

(*Received 30 January 2015, Accepted for publication 10 August 2015*)

Abstract: Sensing of high-definition three-dimensional (3D) sound-space information is of crucial importance for realizing total 3D spatial sound technology. We have proposed a sensing method for 3D sound-space information using symmetrically and densely arranged microphones. This method is called SENZI (Symmetrical object with ENchased Zillion microphones). In the SENZI method, signals recorded by the microphones are simply weighted and summed to synthesize a listener's head-related transfer functions (HRTFs), reflecting the direction in which the listener is facing even after recording. The SENZI method is being developed as a real-time system using a spherical microphone array and field-programmable gate arrays (FPGAs). In the SENZI system, 252 electric condenser microphones (ECMs) were almost uniformly distributed on a rigid sphere. The deviations of the microphone frequency responses were compensated for using the transfer function of the rigid sphere. To avoid the degradation of the accuracy of the synthesized sound space by microphone internal noise, particularly in the low-frequency region, we analyzed the effect of the signal-to-noise ratio (SNR) of microphones on the accuracy of synthesized sound-space information by controlling condition numbers of matrix constructed from transfer functions. On the basis of the results of these analyses, a compact SENZI system was implemented. Results of experiments indicated that 3D sound-space information was well expressed using the system.

Keywords: Sound field recording, Head-related transfer function (HRTF), Spherical microphone array, Electric condenser microphone (ECM), Field-programmable gate array (FPGA)

PACS number: 43.60.Dh, 43.60.Fg, 43.60.Jn, 43.58.Ta [doi:10.1250/ast.36.516]

1. INTRODUCTION

Sensing and reproduction of accurate three-dimensional (3D) sound-space information is of crucial importance for realizing highly realistic audio communications. Tracking of the listener's head motion enhances the reality of captured 3D sound information [1–5]. Therefore, a key to realizing the capture of highly realistic 3D sound information must be the tracking of the (head) motion of a listener who listens to sound captured by the system. In this regard, a few methods have been proposed to sense 3D sound-space information and reproduce it by binaural synthesis [6–9]. In these methods, various microphone arrays are used to capture 3D sound space.

Algazi *et al.* proposed a motion-tracked binaural (MTB) recording technique [6]. In this technique, instead of a dummy head, a sphere or a cylinder with several pairs of microphones is used. The pairs of microphones are installed at opposite positions along the circumference: one pair is selected in accordance with the movement of the listener's head when sound is recorded. This technique was modified, as introduced in their subsequent report [7].

*e-mail: saka@ais.riec.tohoku.ac.jp

In the modified technique, synthesized head-related transfer functions (HRTFs) are personalized by incorporating the shape of the listner's ear, particularly addressing the effect of the pinna. However, synthesized HRTFs are insufficiently personalized because the listener's head size and shape are not considered.

As another approach to sensing or reproducing sound information accurately, a binaural version of High-order Ambisonics (HOA) [10,11] has been developed [8,9]. In this method, a recorded sound space is encoded on several components with specific directivities using spherical harmonic decomposition and is decoded to a virtually arranged loudspeaker array. Then, the signals presented from the virtual loudspeakers are convolved with the listener's HRTFs corresponding to the positions of the virtual loudspeakers. Finally, the synthesized signals are presented binaurally, typically via headphones and transaural systems with a few loudspeakers. Spherical harmonics decomposition is substantially compatible with the spherical shape. Therefore, spherical microphone arrays are used to sense a 3D sound space. However, it remains unclear which order of HOA is sufficient to yield a directional resolution that satisfies the perceptual resolution.

Recently, as a microphone array, a virtual artificial head (VAH) has been developed [12,13]. In the system, the desired frequency-dependent directivity patterns corresponding to the listener's HRTFs are to be resynthesized using a set of spatially distributed microphones with digital filtering. The inter-microphone distance is determined to not match the half-wavelength of the sound from all arrival directions [12]. Filter coefficients are calculated by minimizing a least-squares cost function that connects the desired directivity to the resulting directivity. This technique is applied to various microphone arrays such as a spherical microphone array or a rectangular microphone array. However, the number of microphones used in the developed VAH is only 48. Moreover, these microphones are distributed only in the horizontal area. Therefore, it cannot synthesize accurate sound-space information from all directions in the audible range.

We proposed a method that enables accurate sensing and recording of sound-space information [14–17]. The information recorded by this method can not only be transmitted to distant places in real time but also be stored in media and properly reproduced in the future, even with the appropriate motion of the listener's head tracking at the time of listening. The key hardware component of this method is a microphone array on a human-head-sized rigid sphere with numerous microphones on its surface. The proposed method based on a spherical microphone array is named SENZI (Symmetrical object with ENchased ZIllion microphones), which can be translated as "a thousand ears" in Japanese.

In this study, we introduce the method of implementing the SENZI method in a real-time system using a spherical microphone array and field-programmable gate arrays (FPGAs). In the system, 252 electric condenser microphones (ECMs) were distributed almost uniformly on the human-head-sized rigid sphere. To achieve an accuracy close to that expected when using the SENZI method with 252 microphones, we examined how to reduce the effect of the signal-to-noise ratio (SNR), determined from microphone internal noise and frequency responses, of microphones on the accuracy of synthesized sound-space information. On the basis of the results of these analyses, a compact SENZI system was implemented.

In Sect. 2, an outline of the SENZI method is introduced. According to the algorithm of the SENZI method, the actual 3D sound-space acquisition system using a 252-channel spherical microphone array was developed. In Sect. 3, we present a detailed description of the developed system. Then, in Sects. 4 and 5, we describe the analyses of the effects of the microphone characteristics such as SNR and frequency response on the accuracy of the synthesized sound space. Then, referring to these results, compensation methods for the degradation caused by the microphone characteristics are proposed and the effects of their application to the developed system are described.

2. OUTLINE OF SENZI METHOD [14–17]

The SENZI method comprises a compact human-head-sized solid spherical object with a microphone array on its surface. The microphones are distributed uniformly and symmetrically to accommodate and adapt to the listener's head rotations. Binaural signals are synthesized from the recorded signals and are output to the listener in accordance with the listener's HRTFs.

To calculate and synthesize the listener's HRTFs using inputs from spatially distributed multiple microphones, recorded signals from each microphone are simply weighted and summed to synthesize the listener's HRTF. Let $\boldsymbol{H}_{\mathrm{lis}}$ signify a specified listener's HRTFs for one ear as a function of the direction of the sound source. For a certain frequency f, $\boldsymbol{H}_{\mathrm{lis},f}$ is expressed as follows:

$$\begin{bmatrix} H_{\mathrm{lis},f}(\theta_1) \\ \vdots \\ H_{\mathrm{lis},f}(\theta_m) \end{bmatrix} = \begin{bmatrix} H_{1,f}(\theta_1) & \cdots & H_{n,f}(\theta_1) \\ \vdots & \ddots & \vdots \\ H_{1,f}(\theta_m) & \cdots & H_{n,f}(\theta_m) \end{bmatrix} \begin{bmatrix} z_{1,f} \\ \vdots \\ z_{n,f} \end{bmatrix}. \quad (1)$$

In Eq. (1), $H_{i,f}(\theta_j)$ is the object-related transfer function of the sound propagation path between the i-th microphone and the j-th sound source in the direction of θ_j. $z_{i,f}$ is the weighting coefficient of the i-th microphone at the frequency f. The optimum $\hat{\boldsymbol{z}}_f$ is given by solving Eq. (1) by, for example, the least-mean-squares (LMS) method in

the frequency domain [18]. The result is represented as follows:

$$\boldsymbol{H}_{\mathrm{lis},f} = \boldsymbol{H}_f \cdot \hat{\boldsymbol{z}}_f + \boldsymbol{\varepsilon}, \tag{2}$$
$$\boldsymbol{H}_f = \left[\boldsymbol{H}_{1,f} \ \cdots \ \boldsymbol{H}_{n,f}\right],$$
$$\boldsymbol{H}_{i,f} = \left[H_{i,f}(\theta_1) \ \cdots \ H_{i,f}(\theta_m)\right]^{\mathrm{T}},$$
$$\hat{\boldsymbol{z}}_f = \left[\hat{z}_{1,f} \ \cdots \ \hat{z}_{n,f}\right]^{\mathrm{T}}.$$

Each coefficient $\hat{z}_{i,f}$ is given for each microphone at each frequency. The calculated $\hat{z}_{i,f}$ is a constant complex that is common to all directions of the sound sources. For this reason, the sound source positions need not be considered at all to sense sound-space information coming from any number of sound sources including early reflections and reverberation. This is an important benefit of the proposed method.

For the SENZI method, m transfer functions of all sound propagation paths from all sound sources including reverberation and reflection to the listener's ear are expressed as the weighted sum of $m \times n$ ($m \gg n$) transfer functions of all sound propagation paths from all sound sources to n microphones. The accuracy of the synthesized sound space depends on the number of microphones. Therefore, it is important to use numerous microphones (n) to record the sound space. To accurately synthesize the sound space, it is also important to consider the arrangement of the sound sources for calculating $\hat{z}_i$. Although huge number of the sound sources should be distributed densely and uniformly to cover the entire sound space, reducing the number of sound sources contributes to an improvement in the robustness of the method. Knowledge of human auditory perception can be introduced to reduce the number of sound sources.

However, this SENZI method does not consider the effect of the distances of sound sources. To calculate $\hat{\boldsymbol{z}}_f$ in Eq. (2), only one set of HRTFs is required for one listener. We assume that far-field HRTFs are used to calculate $\hat{\boldsymbol{z}}_f$. This means that the SENZI method cannot accurately reproduce a near-field sound space where HRTFs change depending on the distance. Morimoto *et al.* showed that HRTFs over a 1 m distance are almost independent of the distance of a sound source [19]. Therefore, the SENZI method mainly reproduces a sound space more than about 1 m from the listener.

3. IMPLEMENTATION OF REAL-TIME SENZI SYSTEM [17]

In general, numerous microphones are required to record a 3D sound space accurately. These microphones should be positioned with equal density over the surface of the array to cover the entire 3D space. Although some recording systems have been proposed [12,13,20–22], the accuracy of sound space synthesized by the systems is not sufficient because of the inadequate number of microphones. Moreover, the system size of existing techniques is extremely large.

Fig. 1 Photograph of constructed spherical microphone array.

In contrast, our developed system consists of a spherical microphone array with 252 microphones and FPGAs. These components are sufficiently compact that it is easy to carry all the systems to a distant location for recording. In this section, we explain the details of the developed system.

Figure 1 shows the implemented spherical microphone array. The object radius is 0.085 m. This size is determined on the basis of the average size of a human head. This object, made of epoxy resin, is processed using stereolithography. On the spherical object, 252 microphones are installed. The position of each microphone is calculated on the basis of a regular icosahedron. Each surface of a regular icosahedron is divided into 25 small equilateral triangles. All apices of these triangles are projected to the surface of the spherical object. These 252 points are used as microphone positions. The intervals between all neighboring microphones are almost the same: about 0.02 m. Consequently, the limit for the array's spatial resolution [23] appears at a frequency of more than approximately 8.5 kHz. A small digital omnidirectional electric condenser microphone (ECM) (KUS5147; Hosiden Co., Ltd.) is set at one of the 252 calculated positions. The typical SNR of this microphone is specified as 58 dB (typ.). Recorded signals are 1-bit audio signals at a sampling frequency of 2.4 MHz. These signals are multiplexed and transmitted to the FPGA system through only four wires, which are threaded

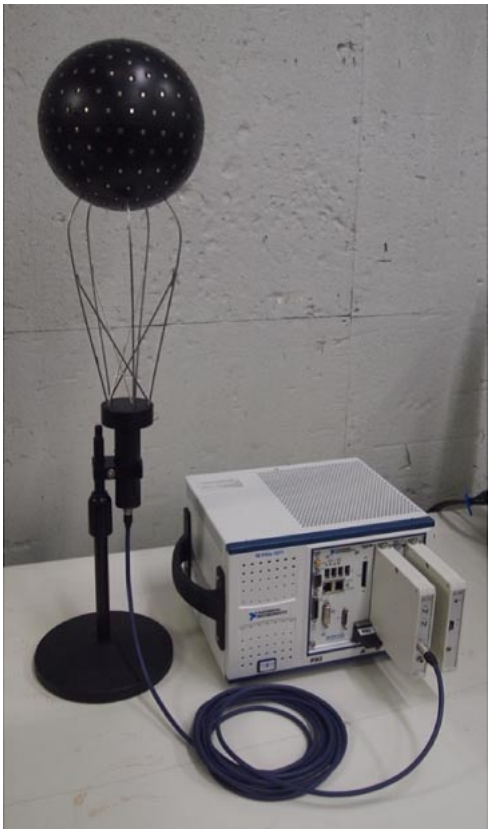


Fig. 2 Actually built SENZI system.

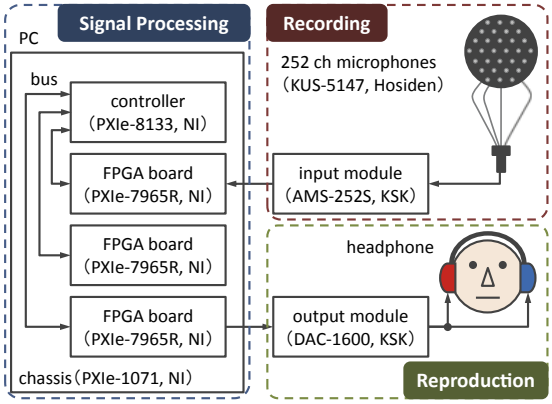


Fig. 3 Component architecture of SENZI system.

through four of five pipes between the microphone array and the support.

Figure 2 shows the constructed real-time SENZI system. This system consists of a "Recording part," a "Signal Processing part," and a "Reproduction part," as shown in Fig. 3. The Recording part simultaneously receives the sound through 252 microphones and sends it to the Signal Processing part after formatting the data ($f_s = 48$ kHz, 16 bit). Then the Signal Processing part processes the 252 input sounds by multiplying the weighting coefficients calculated from the HRTFs of a specified listener. These coefficients are changed in accordance with the head position obtained using the 3D sensor. The Reproduction part generates 2 ch binaural output sound signals provided to the listener, typically through the headphones. The system controller consists of a chassis (PXIe-1071; NI), a controller (PXIe-8133; NI), and an FPGA board (PXIe-7965R; NI). In the system, three FPGA boards are used to operate the 252 sound signals in real time.

Figure 4 shows the process flow of the constructed system on the three FPGA boards. FPGA board 1 is the Recording part. The other boards function as the Signal Processing part and the Reproduction part. The main task of FPGA board 1 is to convert the inputted 252 ch 1-bit audio data into 16-bit data at a sampling frequency of 48 kHz. The converted 252 ch data are transferred to FPGA board 2. On FPGA board 2, the 252 ch data are windowed (Hanning window, 512 points long) and are analyzed using a 512-point FFT with 256-point overlap. Then, on FPGA board 3, binaural signals are calculated and presented to the listener.

The latency of the developed system stems largely from the calculation of Eq. (2). As mentioned above, input signals are analyzed by 512-point FFTs with a 256-point overlap. According to the analysis, the latency was less than 16 ms (512-point FFT + process time).

The length of head-related impulse responses (HRIRs) were decided on the basis of the RIEC HRTF dataset [24]. In the dataset, HRTFs were calculated from measured 512-point HRIRs at a sampling frequency of 48 kHz. Therefore, in our system, we decided to use the 512-point FFT in accordance with the length of the HRIRs.

With the use of these components, our SENZI algorithm is developed as a much more compact system than other proposed systems. This compact nature is an important benefit of our system.

4. EFFECT OF SNR OF MICROPHONES ON SYNTHESIZED-SOUND-SPACE ACCURACY

When the performance of the developed system is analyzed, it is important to evaluate the effect of various noises on synthesized-sound-space information. In this study, we specifically examine the noise related to microphones.

As noise related to the microphones, we consider the following factors: internal noise, microphone position misalignment, variation of microphone frequency responses. It is impossible to compensate for the effect of internal noise perfectly because internal noise varies over time. In contrast, the last two factors are fixed so that it is sufficient to measure these characteristics only once and to compensate for them. Therefore, the effect of first factor should be analyzed separately from those of the last two factors. Moreover, microphone position misalignment can be

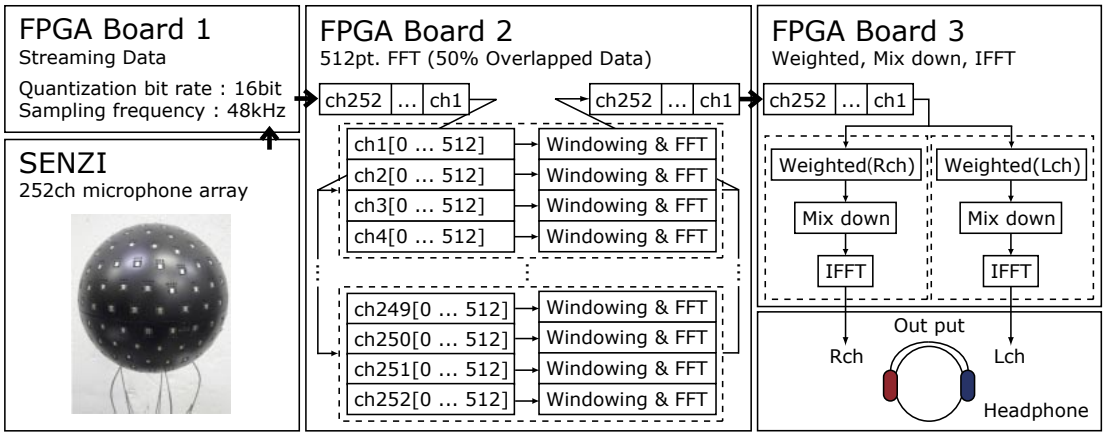


Fig. 4 Process flow of the developed SENZI system.

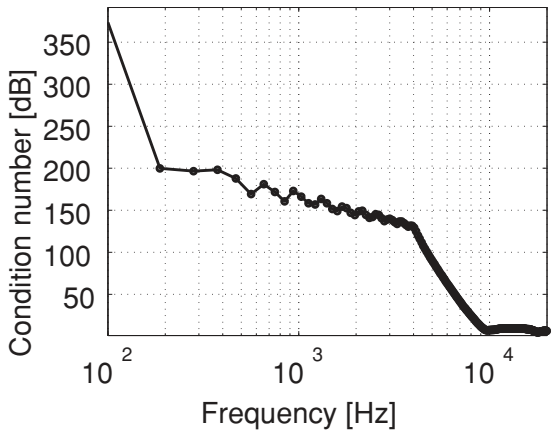


Fig. 5 Condition number of the matrix as a function of frequency.

regarded as the variation of the sensitivity or frequency response of the microphone because only the distance between sound sources and the microphones are changed. Therefore, in the next section, the effect of the variation of microphone frequency responses including sensitivity is focused on. In this section, the effect of internal noise is considered.

In the SENZI system, highly similar signals are recorded from neighboring microphones because the 252 microphones are arranged densely on the sphere. Therefore, when the weighting coefficients are calculated, the matrix may be ill-conditioned, particularly at low frequencies. Figure 5 shows the condition number of the matrix as a function of frequency. This figure indicates that internal noise is expected to affect the accuracy of synthesized-sound-space information more strongly at low frequencies than at high frequencies. As already mentioned above, the average intervals between all neighboring microphones are approximately 0.02 m. This means that the limit for the array's spatial resolution appears at a frequency of more than approximately 8.5 kHz. In Fig. 5, the condition number at a frequency of 8.5 kHz is about 20 dB. Therefore, in this paper, the threshold of rounding off small singular values was set to 20 dB. Although only the results for 20 dB are shown in the paper owing to the limited space, almost the same results were confirmed for other condition numbers. To improve the accuracy of the synthesized sound space, however, the threshold of rounding off small singular values should be optimized at all frequencies. As indicated by the various indexes including a subjective point of view, the effect of the threshold on the accuracy of synthesized sound space should be evaluated in future research.

In this section, we analyzed the effect of the SNR of the microphone on synthesized-sound-space accuracy on the basis of the results of computer simulation and explained how to reduce this effect.

4.1. Simulation

The impulse was assumed to be presented from a certain specified direction $\Omega_1(=(\theta_1,\phi_1))$. θ_1 and ϕ_1 respectively represent the azimuth and elevation angles. The signal $S_f(\Omega_1)$ recorded by the nearest microphone from the sound source was defined as $S_f(\Omega_1)H_{i,f}(\Omega_1)$. As the noise signal generated by the microphone, white noise was added to each recorded signal. Therefore, the signal weighted by the coefficient $\hat{z}_{i,f}$ is expressed as $S_f(\Omega_1)H_{i,f}(\Omega_1)+N_{i,f}$, where $N_{i,f}$ is the internal noise added to the signal recorded by the i-th microphone and n is the total number of microphones.

Using these signals, the synthesized HRTFs of the listener's ear are expressed as

$$\sum_{i=1}^{n}\{S_f(\Omega_1)H_{i,f}(\Omega_1)+N_{i,f}\}\hat{z}_{i,f} = \sum_{i=1}^{n}S_f(\Omega_1)H_{i,f}(\Omega_1)\hat{z}_{i,f}+\sum_{i=1}^{n}N_{i,f}\hat{z}_{i,f}. \quad (3)$$

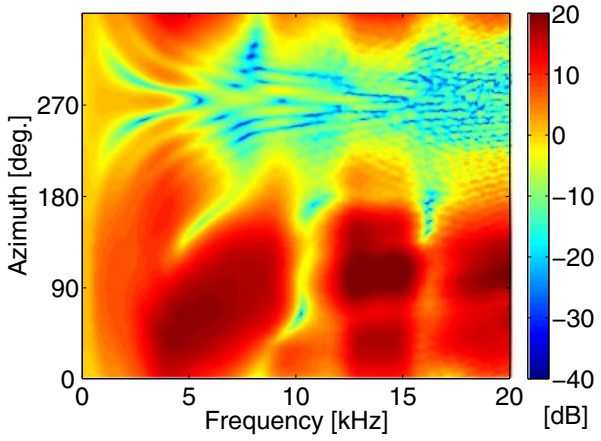


Fig. 6 Target HRTFs of the dummy head (SAMRAI) on the horizontal plane (left ear).

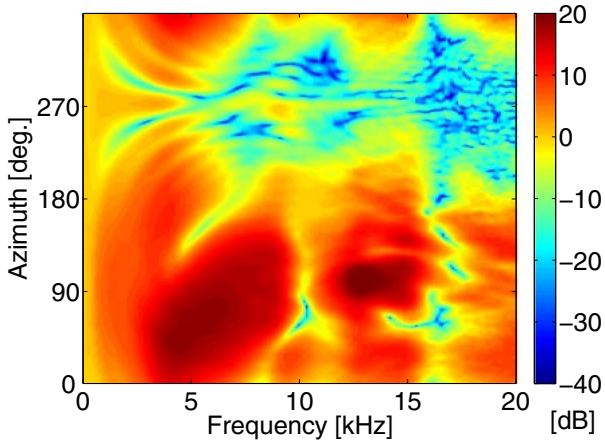


Fig. 7 Ideal HRTFs of the dummy head (SAMRAI) on the horizontal plane (left ear).

Therefore, the effect of the internal noise is expressed as $\sum_{i=1}^{n} N_{i,f}\hat{z}_{i,f}$. Although $N_{i,f}$ is not dependent on the microphones, the expected $N_{i,f}$ would be same. Therefore, the amplitude of each $N_{i,f}$ was set to be the same. The SNR was defined as the ratio of $S_f(\Omega_1)$ to $N_{i,f}$.

The target HRTFs were those of the left ear of a dummy head (SAMRAI; Koken). The 2,562 sound sources were distributed equally on the sphere with a radius of 1.5 m. The 2,562 sound source positions were determined as follows. A regular icosahedron inscribed in a sphere with a radius of 1.5 m was assumed. After that, each surface of the regular icosahedron was divided into 256 small equilateral triangles. All apices of these triangles were projected to the surface of the sphere. Finally, 2,562 apices were obtained at the distance of 1.5 m from the center of the sphere and used as sound source positions. Then, the HRTFs were computed numerically by the boundary element method [25]. Figure 6 shows the target HRTFs on the horizontal plane ($\phi = 0$), which will be referred to hereinafter as target HRTFs. In this figure, 0° is defined as the dummy head's frontal direction and the counterclockwise direction is positive. HRTFs were calculated using the head-related impulse response (HRIR). The length of the HRIR was 512 points at a sampling frequency of 48 kHz.

Transfer functions of sound propagation paths between the positions of 2,562 sound sources and all microphones on the sphere were also calculated analytically [26]. The weighting coefficients $\hat{z}_{i,f}$ expressed in Eq. (2) were calculated using these transfer functions and the HRTFs. Then, target HRTFs were synthesized by the SENZI method for the spherical microphone array with 252 microphones with ideal characteristics (SNR $= \infty$) as the reference. These reference HRTFs were calculated on the horizontal plane from 0° to 359° at 1° steps. The synthesized HRTFs are presented in Fig. 7. These synthesized HRTFs are called ideal HRTFs hereafter.

The amplitude of the signal (S) and the internal noise ($N_{i,f}$) were determined as follows. First, a sound source S was positioned in the direction of Ω. Then, the amplitude of the internal noise ($\boldsymbol{N}$) of the nearest microphone from the sound source was selected at an SNR of 60 dB. The calculated amplitude was applied to that of the internal noise of other microphones. Therefore, the effective SNR is 60 dB or less among the positions of the microphones, except the nearest one, because the amplitude of the sound source decreases with the distance between the sound source and the microphone. These processes were carried out at all sound source positions. The spectral distortion (SD) calculated using Eq. (4) was used as the index of the accuracy of the synthesized sound space.

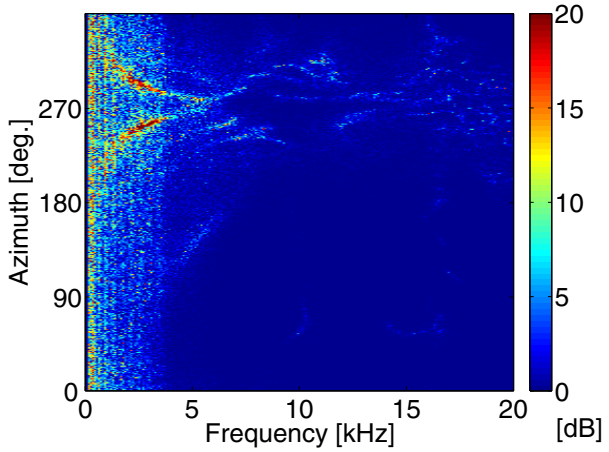
$$\varepsilon_{SD}(f,\theta,\phi) = \left| 20\log_{10}\left| \frac{HRTF_{\text{ideal}}(f,\theta,\phi)}{HRTF_{\text{actual}}(f,\theta,\phi)} \right| \right| \text{ [dB]} \quad (4)$$

θ and ϕ respectively represent the azimuth and elevation angles.

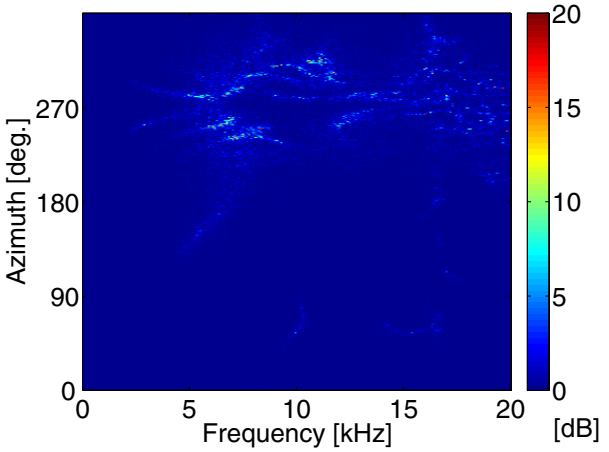
The difference between the target and ideal HRTFs represents the accuracy of the SENZI method itself. Because this difference was already reported previously [16], only the effect of the microphone SNR on the actual HRTFs synthesized by the system is focused on in this paper.

4.2. Results

The results of the simulation are shown in Fig. 8. These figures depict the spectral distortion (SD) of actual HRTFs (SNR = 60 dB) on the horizontal plane. Large synthesized error is observed, especially in the low-frequency regions when the condition number is not considered [Fig. 8(a)]. This error is caused by the SNR of the microphones. Therefore, we introduced signal processing to improve the precision of the calculated HRTFs at an SNR of 60 dB, by



(a) Without consideration of condition number.



(b) With consideration of condition number (Maximum condition number: 20 dB).

Fig. 8 Spectral distortion of actual HRTFs (Simulated SNR of the microphones: 60 dB).

considering the condition numbers. That is, we rounded off small singular values at each frequency to 0 when the condition number is greater than 20 dB [Fig. 8(b)]. By rounding off small singular values, the rank of $\boldsymbol{H}_f$ is decreased. This means that some of the basis vectors are not used to synthesize HRTFs, resulting in better HRTF synthesis performance at low frequencies.

5. EFFECT OF FREQUENCY RESPONSE OF MICROPHONES ON THE ACCURACY OF SYNTHESIZED SOUND SPACE

The implemented system has 252 ch microphones distributed on a rigid sphere. To calculate weighting coefficients, a vast number of transfer functions of the sound propagation paths between the assumed sound positions and all microphones must be used to synthesize HRTFs accurately. The weighting coefficients are calculated using these analytically calculated transfer functions [26]. Therefore, it is important to calibrate the frequency responses of all microphones and reflect them in the recorded signals.

In this section, we introduce the method of measuring frequency responses of all microphones on the sphere. Here, we analyze the effects of the frequency responses of the microphones on the accuracy of the synthesized sound space.

5.1. Measurement of Frequency Responses of 252 Microphones

It is difficult to measure the frequency response of each microphone separately because the relative direction between the loudspeaker and spherical microphone array must be changed accurately. Therefore, the rigid surface is divided into several regions. Then, the frequency responses of the microphones are measured altogether in each region.

When the frequency responses of the microphones are measured in an anechoic room, we assume that (i) relative positions between the loudspeakers and the microphone array are known, and (ii) the loudspeakers can be regarded as a single source from the microphone array. Then, the sound S is presented via the loudspeaker at the position of $\Omega(=(\theta,\phi))$ around the spherical microphone array and recorded by the i-th microphone. The recorded signal $X_{i,f}(\Omega)$ includes the frequency response of the microphone. Therefore, $X_{i,f}(\Omega)$ is expressed as

$$X_{i,f}(\Omega) = S_f(\Omega)H_{i,f}(\Omega)R_{i,f}. \tag{5}$$

In this Eq. (5), $R_{i,f}$ denotes a complex component of the frequency response of the i-th microphone at the frequency of f. This $X_{i,f}(\Omega)$ is given as the product of the sound source $S_f(\Omega)$, object-related transfer function $H_{i,f}(\Omega)$, which is the transfer function of the sound propagation path from the sound source to the i-th microphone, and $R_{i,f}$. Because the shape of the microphone array is a rigid sphere, $\boldsymbol{H}_{i,f}$ can be calculated analytically. Therefore, by dividing $X_i(\Omega)$ into $S(\Omega)$ and $H_i(\Omega)$, R_i is obtained. This Ω for each microphone is selected to calculate the frequency response with a high SNR.

According to Eq. (5), the frequency responses of all microphones were actually measured by the above-described technique. Before that, the frequency response of the loudspeaker (FE83; FOSTEX) was measured using a condenser microphone (4165; B&K). Then, the signal presented from the loudspeaker was compensated for using the inverse characteristic of the frequency response of the loudspeaker. To avoid the effect of background noise, especially in the shadow area, some loudspeakers located around the microphone array should be used. However, because this measurement was conducted in an anechoic room, the effect of the background noise was negligible,

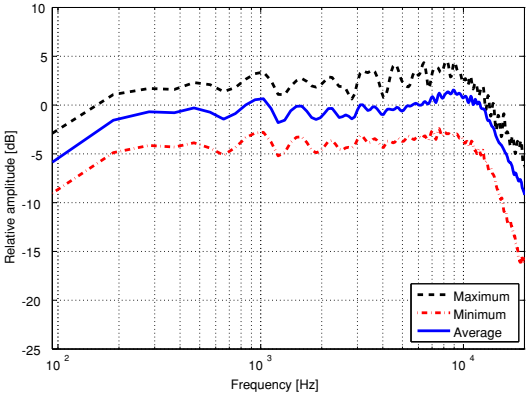


Fig. 9 Magnitudes of frequency responses of all microphones.

even in the shadow area. Therefore, to calculate the frequency responses of all microphones, a loudspeaker was set in front of the microphone array at a distance of 1.5 m and the sound signals produced by the loudspeaker were recorded by all microphones.

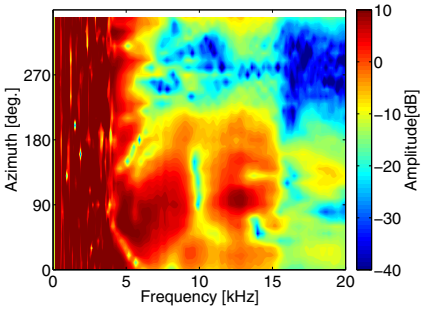
Figure 9 shows the magnitude of measured frequency responses of all microphones. The frequency responses were obtained using by the 512-point FFT of the measured impulse response of the microphones. This figure shows that the frequency responses of all microphones are almost the same, except for the overall gain.

5.2. Evaluation of Accuracy of HRTFs Synthesized by the SENZI System

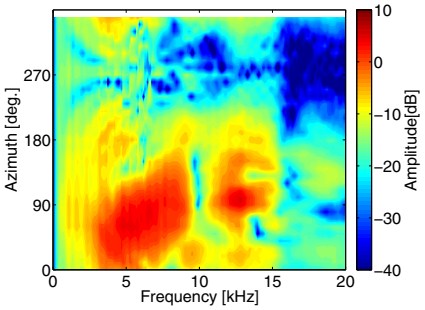
The same target HRTFs [Fig. 6] and ideal HRTFs [Fig. 7] as described in Sect. 4 are used. For the ideal HRTFs, the calculated HRTFs at the step of 10° were extracted and compared with the HRTFs synthesized using measured transfer functions.

Next, the transfer functions of the sound propagation path between the position of the loudspeakers and all microphones were measured. The loudspeaker and spherical microphone array were set on the horizontal plane in an anechoic room. The distance between the loudspeaker and the microphone array was 1.5 m. The loudspeaker position was moved from 0° to 350° in 10° steps. The frequency response of each microphone was compensated with the inverse characteristic of that of the microphone obtained in Sect. 5.1. HRTFs on the horizontal plane were calculated using measured transfer functions and the weighting coefficients that were calculated in Sect. 4. The spectral distortion (SD) calculated by Eq. (4) was used as the index of the accuracy of the synthesized sound space.

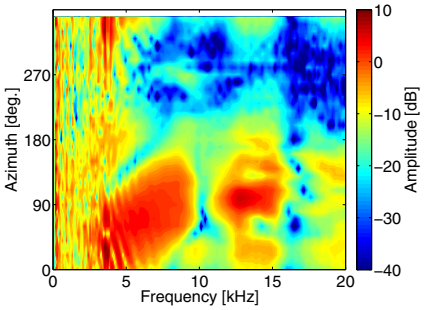
Synthesized HRTFs calculated using measured transfer functions with and without calibration are shown in Fig. 10. As shown in Fig. 10(d), to reduce the degradation of the accuracy in the low-frequency region, a small



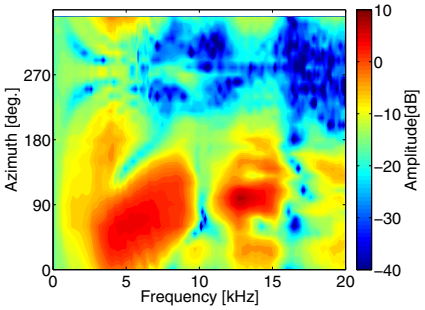
(a) Without calibration.



(b) Without calibration (Maximum condition number: 20 dB).



(c) With calibration.



(d) With calibration (Maximum condition number: 20 dB).

Fig. 10 Actual HRTFs on the horizontal plane synthesized using measured transfer functions with and without calibration of the frequency response of all microphones.

singular value at each frequency was rounded off to 0 when the condition number was greater than 20 dB. Figure 11 shows the spectral distortion between the ideal HRTFs [Fig. 7] and the actual HRTFs synthesized using measured transfer functions [Fig. 10]. As shown in Fig. 9, the frequency response of the microphones is not flat and varies among the microphones. These are the reasons for the degradation of the accuracy of HRTFs synthesized using measured transfer functions without calibration shown in Fig. 11(a). Although the accuracy of the HRTFs is improved by compensating for the frequency responses [Fig. 11(c)], some error remains, especially in the low-frequency region. Such an error can be decreased [Fig. 11(d)] by controlling the condition numbers of the matrix. On the other hand, to obtain accurate sound space information in the high-frequency region, the calibration is effective, as shown in Fig. 11(b). These results show that rounding off the small singular value is effective not only for reducing the effect of microphone internal noise, but also for compensating the frequency responses of the microphones.

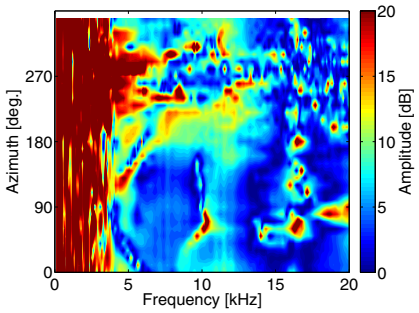
Although the proposed technique seems to work well as a simple way of compensating for the characteristics of the microphones, the effect of the characteristics of the loudspeakers should be considered when more accurate frequency responses of the microphones are required. Such analysis should be performed from the perceptual point of view, as well. Evaluations of the total accuracy of synthesized 3D sound-space information shall be undertaken in future research.

6. CONCLUSION

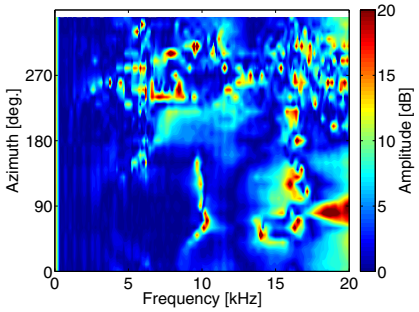
In this study, the developed SENZI system with a human-head-sized solid spherical microphone array and field-programmable gate arrays (FPGAs) was introduced. The system consists of three FPGA boards and a 252 ch spherical microphone array. Electric condenser microphones (ECMs) were distributed almost uniformly over the rigid surface.

To avoid degradation of the accuracy of the synthesized sound space by internal noise, particularly in the low-frequency region, we reduced the effect of the signal-to-noise ratio (SNR) of microphones on the accuracy of synthesized sound-space information by controlling condition numbers of the matrix constructed from transfer functions between the spherical microphone array and sound source positions. The deviations of the microphone frequency responses were also compensated for using the transfer function of the rigid sphere.

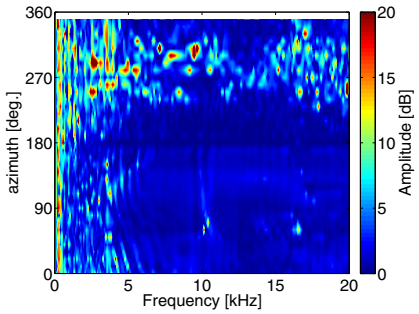
On the basis of these analyses, a compact SENZI system was implemented. The results of experiments show that three-dimensional (3D) sound-space information is well expressed using the system. It is also important to



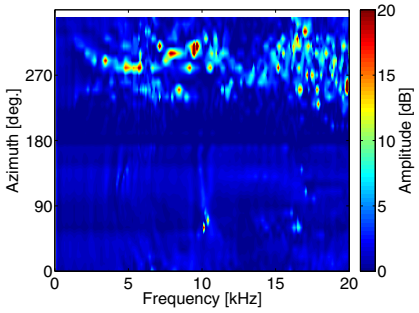
(a) Without calibration.



(b) Without calibration (Maximum condition number: 20 dB).



(c) With calibration.



(d) With calibration (Maximum condition number: 20 dB).

Fig. 11 Spectral distortion of actual HRTFs synthesized using measured transfer functions with and without calibration of the frequency response of all microphones.

evaluate how synthesized 3D sound-space information is perceptually accurate. A detailed investigation about this point shall be undertaken in future research.

ACKNOWLEDGEMENTS

Portions of this report were presented at Acoustics 2012. A part of this work was supported by a grant from the Strategic Information and Communications R&D Promotion Programme (SCOPE) No. 082102005 from the Ministry of Internal Affairs and Communications (MIC), Japan and the A3 Foresight Program for "Ultra-realistic acoustic interactive communication on next-generation Internet." We wish to thank Mr. Ju'nichi Kodama, Mr. Jumpei Matsunaga and Mr. Yuto Wada for their help with analysis of the results.

REFERENCES

[1] H. Wallach, "On sound localization," *J. Acoust. Soc. Am.*, **10**, 270–274 (1939).

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

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

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

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

[6] V. R. Algazi, R. O. Duda and D. M. Thompson, "Motion-Tracked Binaural Sound," *J. Audio Eng. Soc.*, **52**, 1142–1156 (2004).

[7] J. B. Melick, V. R. Algazi, R. O. Duda and D. M. Thompson, "Customization for personalized rendering of motion-tracked binaural sound," *Proc. AES Convention 117*, 6225, http://www.aes.org/e-lib/browse.cfm?elib=12882 (2004).

[8] M. Noisternig, A. Sontacchi, T. Musil and R. Höldrich, "A 3D ambisonic based binaural sound reproduction system," *Proc. AES 24th Int. Conf.*, 1, http://www.aes.org/e-lib/browse.cfm?elib=12314 (2003).

[9] L. S. Davis, R. Duraiswami, E. Grassi, N. A. Gumerov, Z. Li and D. N. Zotkin, "High order spatial audio capture and its binaural head-tracked playback over headphones with HRTF cues," *Proc. AES Convention 119*, 6540, http://www.aes.org/e-lib/browse.cfm?elib=13369 (2005).

[10] M. A. Gerzon, "Periphony: with-height sound reproduction," *J. Audio Eng. Soc.*, **21**, 2–10 (1973).

[11] J. Daniel, "Spatial sound encoding including near field effect: introducing distance coding filters and a viable, new ambisonic format," *Proc. AES 23rd Int. Conf.*, **16**, http://www.aes.org/e-lib/browse.cfm?elib=12321 (2003).

[12] E. Rasumow, M. Blau, M. Hansen, S. Doclo, S. van de Par and V. Mellert, "Robustness of virtual artificial head topologies with respect to microphone positioning errors," *Proc. Forum Acusticum 2011*, pp. 2251–2256 (2011).

[13] E. Rasumow, M. Blau, M. Hansen, S. Doclo, S. van de Par and V. Mellert, "The impact of the white noise gain (WNG) of a virtual artificial head on the appraisal of binaural sound reproduction," *Proc. EAA Jt. Symp. Auralization and Ambisonics*, pp. 174–180 (2014).

[14] R. Kadoi, S. Sakamoto, S. Hongo and Y. Suzuki, "A new recording and reproduction method of sound space information by a microphone array," *Proc. Spring Meet. Acoust. Soc. Jpn.*, 3-Q-18, pp. 685–686 (2007) (in Japanese).

[15] S. Sakamoto, R. Kadoi, S. Hongo and Y. Suzuki, "SENZI and ASURA: New high-precision sound-space sensing systems based on symmetrically arranged numerous microphones," *Proc. 2nd Int. Symp. Universal Communication* (*ISUC 2008*), pp. 429–434 (2008).

[16] S. Sakamto, S. Hongo and Y. Suzuki, "3D sound-space sensing method based on numerous symmetrically arranged microphones," *IEICE. Trans. Fundum.*, **E97-A**, 1893–1901 (2014).

[17] S. Sakamoto, "SENZI: 3D sound-space recording and reproduction method based on spherical microphone array," *J. Acoust. Soc. Jpn.* (*J*), **70**, 379–384 (2014) (in Japanese).

[18] O. Kirkeby and P. Nelson, "Reproduction of plane wave sound fields," *J. Acoust. Soc. Am.*, **94**, 2992–3000 (1993).

[19] M. Morimoto, N. Joren, Y. Ando and Z. Maekawa, "On head-related transfer function," *Trans. Tech. Comm. Psychol. Physiol. Acoust., Acoust. Soc. Jpn.*, H-31-1-2, pp. 4–9 (1976) (in Japanese).

[20] S. Bertet, J. Daniel and S. Moreau, "3D sound field recording with higher order ambisonics — Objective measurements and validation of spherical microphone," *Proc. AES Convention 120*, 6857, http://www.aes.org/e-lib/browse.cfm?elib=13661 (2006).

[21] D. N. Zotkin, R. Duraiswami and N. A. Gumerov, "Plane-wave decomposition of acoustical scenes via spherical and cylindrical microphone array," *IEEE Trans. Audio Speech Lang. Process.*, **18**, 2–16 (2010).

[22] C. Jin, N. Epain and A. Parthy, "Design, optimization and evaluation of a dual-radius spherical microphone array," *IEEE/ACM Trans. Audio Speech Lang. Process.*, **22**, 193–204 (2014).

[23] S. U. Pillai, *Array Signal Processing* (Springer-Verlag, New York, 1989).

[24] K. Watanabe, Y. Iwaya, Y. Suzuki, S. Takane and S. Sato, "Dataset of head-related transfer functions measured with a circular loudspeaker array," *Acoust. Sci. & Tech.*, **35**, 159–165 (2014).

[25] M. Otani and S. Ise, "Fast calculation system specialized for head-related transfer function based on boundary element method," *J. Acoust. Soc. Am.*, **119**, 2589–2598 (2006).

[26] R. O. Duda and W. L. Martens, "Range dependence of the response of a spherical head model," *J. Acoust. Soc. Am.*, **104**, 3048–3058 (1998).

Shuichi Sakamoto received B.S., M.S. and Ph.D. degrees from Tohoku University, in 1995, 1997, and 2004, respectively. He is currently an Associate Professor at the Research Institute of Electrical Communication, Tohoku University. He was a Visiting Researcher at McGill University, Montreal, Canada during 2007–2008. His research interests include human multi-sensory information processing including hearing, speech perception, and development of high-definition 3D audio recording systems. He is a member of ASJ, IEICE, VRSJ, and others.

Satoshi Hongo received B.E., M.E. and Dr. Eng. degrees from Tohoku University in 1990, 1992, and 1995, respectively. From 1995 to 1999, he was a Research Associate with the Faculty of Engineering, Tohoku University. From 1999 to 2008 he was an Associate Professor of Department of Design and Computer Applications, Miyagi National College of Technology (SNCT), Japan. He has been a Professor of the Advanced Course, Sendai National College of Technology and Chairman of the Department of Design and Computer Applications, SNCT.

Takuma Okamoto received a Ph.D. in information sciences in 2009 from Tohoku University, Japan. From 2009, he was a postdoctoral research fellow at the Research Institute of Electrical Communication, Tohoku University, Japan. Since 2012, he is currently a researcher at the Universal Communication Research Institute of the National Institute of Information and Communications Technology, Japan. His research interests include microphone array signal processing, 3D sound field recording and reproduction, and active sound field control. From 2014, he extended his research area to automatic speech recognition, speech synthesis and spoken dialogue systems. He received the Awaya Prize Young Researcher Award from the Acoustical Society of Japan in 2012.

Yukio Iwaya graduated from Tohoku University in 1991 and received his Ph.D. degree in information sciences in 1999. He is currently a Professor of Tohoku Gakuin University. His research interests include three-dimensional acoustic space perception and development of its communication systems with high sense of presence.

Yôiti Suzuki graduated from Tohoku University in 1976 and received his Ph.D. degree in electrical and communication engineering in 1981. He is currently a Professor of Research Institute of Electrical Communication, Tohoku University. His research interests include psychoacoustics and digital signal processing of acoustic signals. He served as a president of the Acoustical Society of Japan from '05 to '07. He is a fellow of the Acoustical Society of America.