24th INTERNATIONAL CONGRESS ON SOUND AND VIBRATION
23–27 July 2017, London

ICSV24
23-27JULY 2017
LONDON CALLING

BINAURAL RENDERING OF SPHERICAL MICROPHONE ARRAY RECORDINGS BY DIRECTLY SYNTHESIZING THE SPATIAL PATTERN OF THE HEAD-RELATED TRANSFER FUNCTION

Shuichi Sakamoto, César Salvador, Jorge Treviño and Yôiti Suzuki
Tohoku University, Research Institute of Electrical Communication and Graduate School of Information Sciences, 2-1-1 Katahira, Aoba-ku, Sendai, 980-8577, Japan
email: saka@ais.riec.tohoku.ac.jp

Binaural technologies can convey rich spatial auditory information to listeners using simple equipment such as headphones. Advanced binaural recording and reproduction methods use spherical microphone arrays and head-related transfer function (HRTF) datasets. Mainstream techniques, such as binaural Ambisonics, characterize the recorded sound field as a weighted sum of spherical harmonics functions. In contrast, this research seeks to generate individualized binaural signals directly from the microphone recordings, without relying on intermediate sound field representations. The approach, known as SENZI, applies a set of weighting filters to the recorded microphone signals resulting in the target spatial pattern defined by the HRTF dataset. In this sense, the proposal requires finding the appropriate weighting filters by inverting a linear system. Binaural synthesis methods based on the solution to an inverse problem belong to one of two categories: HRTF modeling (type 1) or microphone signal modeling (type 2). The SENZI method considered here belongs to the HRTF modeling category. In addition, the problem is generally over- or underdetermined, depending on the number of microphones in the array and HRTFs in the dataset. This also impacts the accuracy of the synthesized binaural signals. A design problem, therefore, is to choose the most appropriate number of microphones and HRTFs. Fortunately, large HRTF datasets, as well as massively multi-channel arrays are now available. An example of the latter is a real-time implementation of the SENZI method using a 252-channel spherical microphone array and a FPGA-based processing subsystem. This research evaluates the binaural synthesis accuracy in relation to the number of microphones and HRTFs used to derive the weighting filters. Numerical simulations show that underdetermined systems generally yield better results than overdetermined ones.

Keywords: 3D audio technology, binaural synthesis, spherical acoustics, head-related transfer functions, and microphone arrays

1. Introduction

Advanced multimedia systems require recording and reproduction technologies that can accurately convey three-dimensional (3D) sound space information to distant places or store it for future access. One of the promising methods to realize this is binaural technologies. These technologies can convey rich spatial auditory information to listeners using simple equipment such as headphones.

One of the important components of the technologies is the head-related transfer function (HRTF). HRTFs contain individual auditory cues that are important to present sounds which are accurately

located outside of the head [1]. Recent advances in the measurement and computation of HRTFs led to public databased for their values in densely sampled spatial grids [2, 3]. The acquisition subsystem is also crucially important component to realize the technologies. This subsystem typically applies microphone arrays. Use of multichannel arrays is essential to sense accurate sound space. Spherical microphone arrays with numerous microphones are typically used because they enable us to analyze the sound space easily using spherical harmonic decomposition.

To design binaural systems with the above mentioned two components, spherical microphone arrays with numerous microphones are typically used because they enable us to flexibly design high definition sound acquisition systems. Such a class of technologies can be classified into two types of general formulations [4]. A class of signal processing method to design such binaural systems relies on the spherical harmonic decomposition to analyze the sound space and has been intensively studied because of its simple and elegant theory [5, 6, 7, 8, 9]. We have been devoting in the other class of signal processing methods that generate individualized binaural signals directly from the microphone recordings, without relying on intermediate sound field representations such as by spherical harmonic decomposition. The approach, known as SENZI [10, 11], applies a set of weighting filters to the recorded microphone signals resulting in the target spatial pattern defined by the HRTF dataset. To realize this approach as an actual system, it is an important issue to choose the most appropriate number of microphones and HRTFs. This is a common issue to develop not only SENZI but also all other binaural technologies using the spherical microphone array and the HRTFs [4].

In this study, we summarize the way of developing these binaural technologies and discuss the relationship between the number of HRTFs and microphones and the point of accuracy of synthesized sound space. According to the analysis, the characteristics of SENZI are discussed to assess its advantages.

2. Binaural synthesis from microphone array recordings and HRTF datasets [4]

An overview of the binaural synthesis method under consideration is illustrated in Fig. 1. In general, this class of binaural synthesis methods aim at rendering a sound pressure field sampled at the positions where a set of HRTFs is given. In other words, this class of methods aim to synthesize the binaural signals due to an array of virtual loudspeakers placed at the positions used to obtain the HRTFs.

In this system, binaural signals to the individual listeners are generated by using HRTFs h_ℓ^{left} and h_ℓ^{right} ($\ell = 1, 2, ..., L$) and recorded signals p_m ($m = 1, 2, ..., M$) as drawn in Fig. 1. It is argued that most of the existing methods for combining the microphone array recordings with HRTF datasets can be classified into two prominent approaches: 1) the HRTF modeling approach and 2) the microphone signal modeling approach. Each approach refers to the manner in which the underlying inverse problem is formulated.

In the HRTF modeling approach, an HRTF dataset constitutes a specified spatial pattern to be approximated by a set of weighting filters $\boldsymbol{w}$ applied to the microphone recordings. $\boldsymbol{w}$ are calculated by solving a linear system of equations that approximate the HRTF dataset. The entries of the matrix associated to the linear system are acoustic transfer functions from the positions of microphones to the positions used to obtain the HRTF dataset. Examples of binaural systems based on this approach are the virtual artificial head (VAH) [12, 13], binaural beamforming systems [14, 15, 16, 17, 18].

In the microphone signal modeling approach, on the other hand, the microphone array recordings are used to calculate the driving signals $\boldsymbol{u}$ at the positions used to obtain the HRTF dataset. $\boldsymbol{u}$ are subsequently rendered with the corresponding HRTFs in the dataset by relying on the principle of acoustic wave superposition [19]. These $\boldsymbol{u}$ are calculated by solving a linear system of equations that model the microphone array recordings. The entries of the associated matrix are acoustic transfer functions from the positions involved in the HRTF dataset to the positions of microphones. Examples

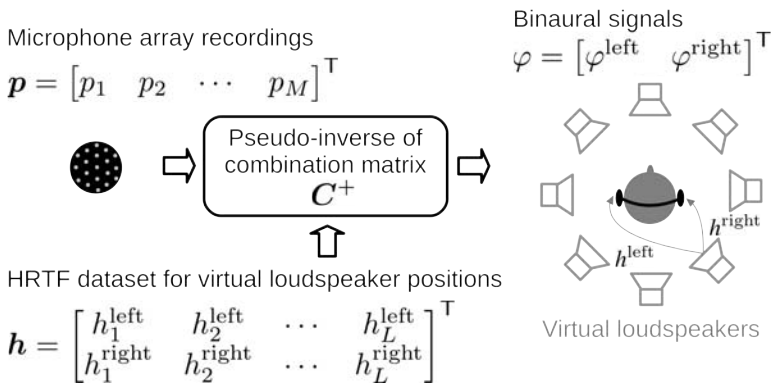


Figure 1: Overview of binaural systems.

of binaural systems based on this approach are the binaural system obtained when combining the BPLIC [20] and ADVISE systems [21], and the binaural ambisonic systems treated in [5, 6, 7, 8, 9].

Generalized binaural synthesis explained previously can be formulated in terms of the following equation:

$$\varphi = \overline{\boldsymbol{h}^\top} \boldsymbol{C}^+ \boldsymbol{p}. \tag{1}$$

Here, the synthesized binaural signals for the left and right ears are organized in

$$\varphi = \begin{bmatrix} \varphi^{\text{left}} & \varphi^{\text{right}} \end{bmatrix}^\top. \tag{2}$$

The symbol $^\top$ indicates transpose and the overbar symbol is used to denote the complex conjugate. $\boldsymbol{C}^+$ is pseudo-inverse of a combination matrix between microphone positions and the virtual loudspeaker positions, denoted by $\boldsymbol{C}$. The recordings of an array composed of M microphones are organized in the vector

$$\boldsymbol{p} = \begin{bmatrix} p_1 & p_2 & \cdots & p_M \end{bmatrix}^\top, \tag{3}$$

while the HRTFs of the dataset are organized in the matrix

$$\boldsymbol{h} = \begin{bmatrix} \boldsymbol{h}^{\text{left}} \\ \boldsymbol{h}^{\text{right}} \end{bmatrix}^\top = \begin{bmatrix} h_1^{\text{left}} & h_2^{\text{left}} & \cdots & h_L^{\text{left}} \\ h_1^{\text{right}} & h_2^{\text{right}} & \cdots & h_L^{\text{right}} \end{bmatrix}^\top. \tag{4}$$

The conditions to calculate $\boldsymbol{C}^+$ are investigated below for each binaural synthesis approach. A combination matrix $\boldsymbol{C}_{\text{HRTF}}$ of size $L \times M$ will be used in the HRTF modeling approach, whereas a combination matrix $\boldsymbol{C}_{\text{mic}}$ of size $M \times L$ will be used in the microphone signal modeling approach. The structures for spatial signal processing that result from these two approaches are illustrated in Fig. 2.

The binaural signals calculated by HRTF modeling approach are synthesized by

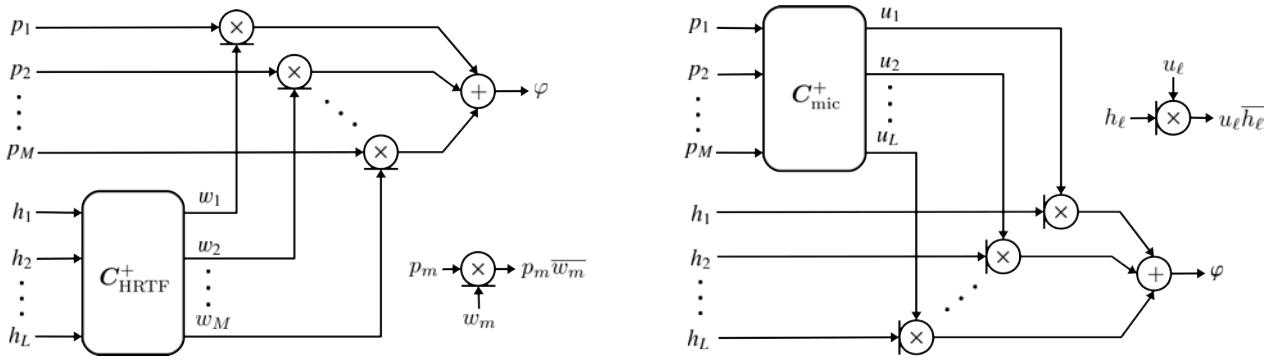
$$\varphi = \overline{\boldsymbol{w}^\top} \boldsymbol{p}, \tag{5}$$

where

$$\boldsymbol{w} = \boldsymbol{C}_{\text{HRTF}}^+ \boldsymbol{h} \tag{6}$$

defines the weighting filters. On the other hand, according to the microphone signal modeling approach, the binaural signals are synthesized by

$$\varphi = \overline{\boldsymbol{h}^\top} \boldsymbol{u}, \tag{7}$$



(a) HRTF modeling.

(b) Microphone signal modeling.

Figure 2: Two approaches for binaural synthesis that combine microphone array recordings (p_m) and HRTF datasets (h_ℓ).

where

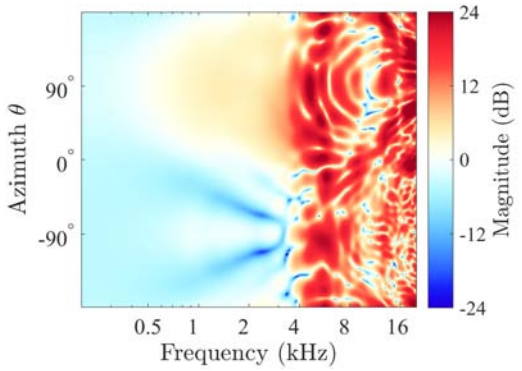
$$\boldsymbol{u} = \boldsymbol{C}^+_{\mathrm{mic}}\boldsymbol{p} \tag{8}$$

defines the driving signals. A remarkable relation between the two synthesis approaches is further verified when comparing the dimensions of the corresponding linear systems. When one approach arises from the solution to an overdetermined system of equations, the other corresponds to an underdetermined system of equations, and vice versa. The synthesized binaural signals using these two modelings are shown in Fig. 3. When the number of microphones is smaller than that of HRTF dataset ($M < L$), $\boldsymbol{C}_{\mathrm{HRTF}}$ represents an overdetermined system, whereas $\boldsymbol{C}_{\mathrm{mic}}$ represents an underdetermined system. In this case, more distortion is observed in the HRTF modeling approach. In contrast, when the nuber of microphones is larger than that of HRTF dataset ($M > L$), $\boldsymbol{C}_{\mathrm{HRTF}}$ corresponds to an underdetermined system, whereas $\boldsymbol{C}_{\mathrm{mic}}$ corresponds to an overdetermined system. In this situation, more distortion is observed in the microphone signal modeling approach.

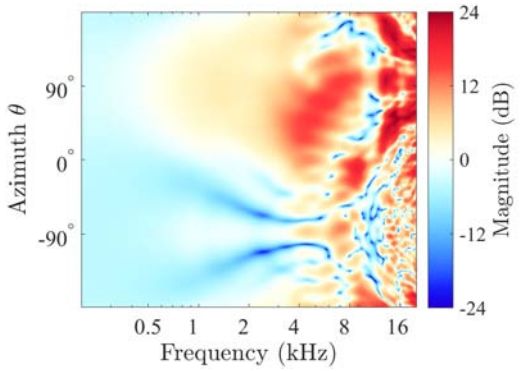
3. Sound space recording and reproduction method, SENZI [22, 11, 23]

Our developed method, SENZI (Symmetrical object with ENchased ZIllion microphones), is one of the methods of HRTF modeling approach. To calculate and synthesize a listener's HRTFs by inputs from spatially distributed multiple microphones, recorded signals from each microphone are simply weighted and summed to synthesize a listener's HRTF. The weight can be changed according to a listener's 3D head movement. Because of that feature, 3D sound-space information is acquired accurately irrespective of the head movement. 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.

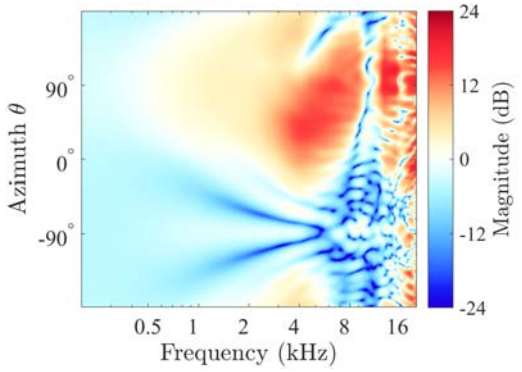
Figure 4 shows the constructed real-time SENZI system. In the system, three FPGA boards are used to operate the 252 sound signals in real time. This system consists of a "Recording part", a "Signal Processing part", and a "Reproduction part." FPGA board 1 is the Recording part. The main task of the FPGA board 1 is to convert the inputted 252 ch 1-bit audio data into 16-bit data at a sampling frequency of 48 kHz. This converted signals send to the Signal Processing part implemented on FPGA board 2 and 3. 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, the 252 input sounds are multiplied by the weighting coefficients calculated from the HRTFs of a specified listener. These coefficients are changed in accordance with the head position obtained using



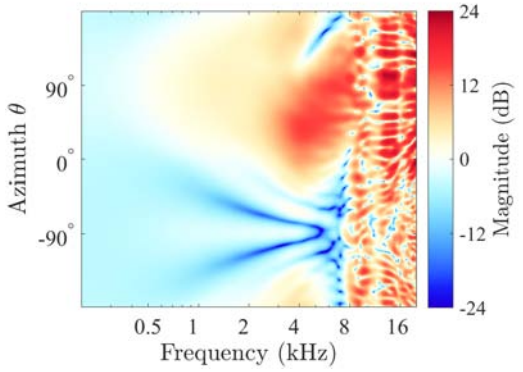
(a) HRTF modeling, $M = 92$, $L = 252$.



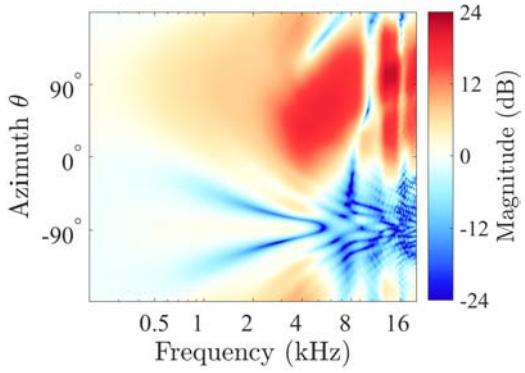
(b) Microphone signal modeling, $M = 92$, $L = 252$.



(c) HRTF modeling, $M = 492$, $L = 252$.



(d) Microphone signal modeling, $M = 492$, $L = 252$.



(e) Reference binaural signals.

Figure 3: Binaural synthesis examples.

the 3D sensor. Finally, the Reproduction part on the FPGA board 3 generates 2 ch binaural output sound signals provided to the listener, typically through the headphones.

Spherical microphone array shown in Fig. 4 is at the radius of 8.5 cm. On the spherical object, 252 microphones are installed. The position of each microphone is calculated on the basis of a regular icosahedron. The intervals between all neighboring microphones are almost the same: average interval is 2.043 cm. Consequently, the limit for the array's spatial resolution appears at a frequency

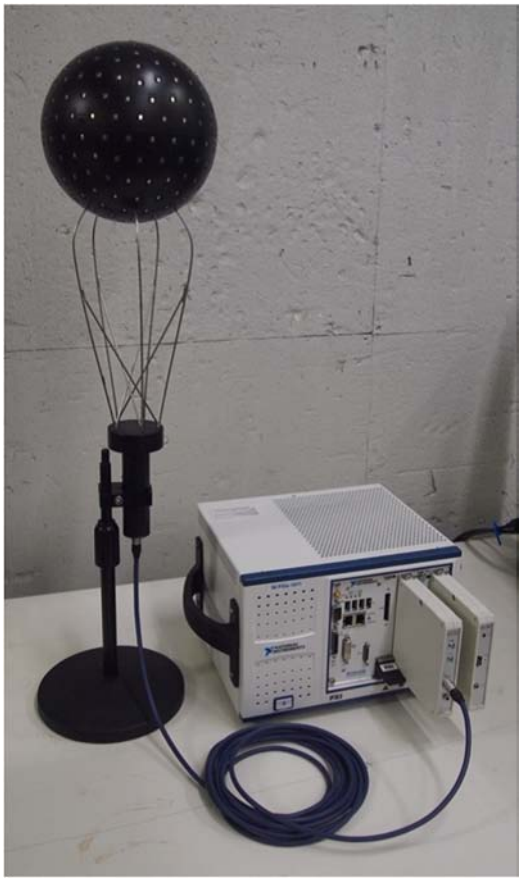


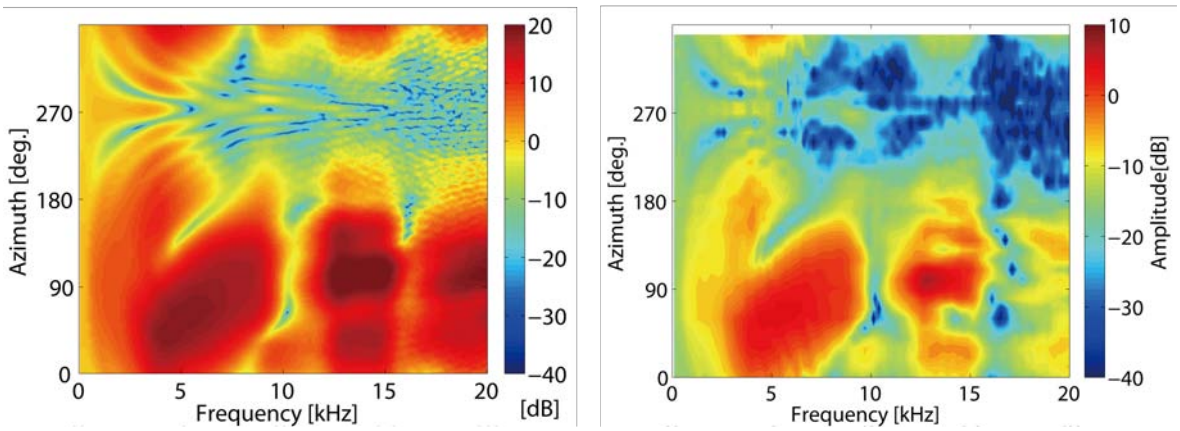
Figure 4: Photograph of SENZI system.

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.).

Synthesized HRTFs using the constructed SENZI system are shown in Fig. 5. To reduce the degradation of the accuracy in the low-frequency region caused by the low signal-to-noise ratio (SNR) of the microphone, a small singular value at each frequency was rounded off to 0 when the condition number was greater than 20 dB. Fig. 5(b) indicates that the dynamic range of synthesized HRTFs is narrower than that of the target HRTFs. This narrow dynamic range is caused by the low SNR of the microphone. Except for this, the frequency characteristics up to around 10 kHz, which is the limit of spatial resolution, are almost well synthesized even when HRTFs are synthesizes using actually recorded signals.

4. Conclusions

It is crucially important to convey accurate auditory spatial information to listeners for development of advanced multimedia communications systems. Binaural technologies are promising methods which can convey rich spatial auditory information to listeners using simple equipment such as headphones. The main components of recent binaural technologies are spherical microphone arrays and HRTFs datasets. In this study, two types of general formulations for such a class of technologies using these components were presented. According to the classification, we explained our proposed binaural synthesis method named SENZI, which is one of the HRTF modeling approaches. Numerical simulations showed that binaural synthesis accuracy of these systems depended on the number of microphones compared to that of HRTFs. More specifically, underdetermined systems yield better



(a) Target HRTFs.

(b) Synthesized HRTFs.

Figure 5: Actual HRTFs on the horizontal plane synthesized using SENZI system.

binaural synthesis accuracy than overdetermined ones.

Acknowledgments

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, the A3 Foresight Program for “Ultra-realistic acoustic interactive communication on next-generation Internet,” and JSPS KAKENHI Grant Numbers JP24240016, JP26280067, JP16H01736.

REFERENCES

1. M. Morimoto and Y. Ando, “On the simulation of sound localization,” *J. Acoust. Soc. Jpn. (E)*, 1(3), 167–174, (1980).

2. P. Majdak, Y. Iwaya, T. Carpentier, R. Nicol, M. Parmentier, A. Roginska, Y. Suzuki, K. Watanabe, H. Wierstorf, H. Ziegelwanger, and M. Noisternig, “Spatially oriented format for acoustics: a data exchange format representing head-related transfer functions,” *AES 134 Convention*, (2013).

3. 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(3), 159–165, (2014).

4. C. Salvador, S. Sakamoto, J. Treviño, and Y. Suzuki, “Design theory for binaural synthesis: combining microphone array recordings and head-related transfer function datasets,” *Acoust. Sci. & Tech.*, 38(2), 51–62, (2017).

5. R. Duraiswami, D. N. Zotkin, Z. Li, E. Grassi, N. A. Gumerov, and L. S. Davis, “High order spatial audio capture and its binaural head-tracked playback over headphones with HRTF cues,” *AES 119*, New York, USA, (2005).

6. A. Avni, J. Ahrens, M. Geier, S. Spors, H. Wierstorf, and B. Rafaely, “Spatial perception of sound fields recorded by spherical microphone arrays with varying spatial resolution,” *J. Acoust. Soc. Am.*, 133(5), 2711–2721, (2013).

7. C. D. Salvador, S. Sakamoto, J. Treviño, and Y. Suzuki, “Numerical Evaluation of Binaural Synthesis from Rigid Spherical Microphone Array Recordings,” *AES Int. Conf. Headphone Technology*, (2016).

8. C. D. Salvador, S. Sakamoto, J. Treviño, and Y. Suzuki, "Spatial accuracy of binaural synthesis from rigid spherical microphone array recordings," *Acoust. Sci. & Tech.*, 38(1), 23–30, (2017).

9. B. Bernschutz, A. V. Giner, C. Porschmann, and J. Arend, "Binaural reproduction of plane waves with reduced modal order," *Acta Acust. United Ac.*, 100(5), 972–983, (2014).

10. R. Kadoi, S. Sakamoto, S. Hongo, and Y. Suzuki, "Sound space information reproduction system by using artificial head model with microphone array," *11th the Virtual Reality Society of Japan Annual Conference*, 127–130, (2006). (in Japanese)

11. S. Sakamoto, S. Hongo, and Y. Suzuki, "3d sound-space sensing method based on numerous symmetrically arranged microphones," *IEICE Trans. Fundamentals*, vol. E97-A, no. 9, 1893–1901, (2014).

12. E. Rasumow, M. Blau, M. Hansen, S. van de Par, S. Doclo, V. Mellert, and D. Püschel, "Smoothing individual head-related transfer functions in the frequency and spatial domains," *J. Acoust. Soc. Am.*, 135(4), 2012–2025, (2014).

13. E. Rasumow, M. Hansen, S. v. d. Par, D. Püschel, V. Mellert, S. Doclo, and M. Blau, "Regularization approaches for synthesizing HRTF directivity patterns," *IEEE/ACM Trans. Audio, Speech, Language Process.*, 24(2), 215–225, (2016).

14. W. Song, W. Ellermeier, and J. Hald, "Using beamforming and binaural synthesis for the psychoacoustical evaluation of target sources in noise," *J. Acoust. Soc. Am.*, 123(2), 910–924, (2008).

15. C. D. Salvador, S. Sakamoto, J. Treviño, J. Li, Y. Yan, and Y. Suzuki, "Accuracy of head-related transfer functions synthesized with spherical microphone arrays," *Proc. Mtgs. Acoust.*, 19(1), (2013).

16. N. R. Shabtai and B. Rafaely, "Generalized spherical array beamforming for binaural speech reproduction," *IEEE/ACM Trans. Audio, Speech, Language Process.*, 22(1), 238–247, (2014).

17. N. R. Shabtai, "Optimization of the directivity in binaural sound reproduction beamforming," *J. Acoust. Soc. Am.*, 138(5), 3118–3128, (2015).

18. C. D. Salvador, S. Sakamoto, J. Treviño, and Y. Suzuki, "Evaluation of white noise gain in a binaural system for microphone arrays," *Acoust. Soc. of Japan autumn meeting*, 3-7-12, 401–404, (2016).

19. G. H. Koopmann, L. Song, and J. B. Fahnline, "A method for computing acoustic fields based on the principle of wave superposition," *J. Acoust. Soc. Am.*, 86(6), 2433–2438, (1989).

20. S. Takane, Y. Suzuki, and T. Sone, "A new method for global sound field reproduction based on Kirchhoff's integral equation," *Acta Acust. United Ac.*, 85(2), 250–257, (1999).

21. S. Takane, Y. Suzuki, T. Miyajima, and T. Sone, "A new theory for high definition virtual acoustic display named ADVISE," *Acoust. Sci. & Tech.*, 24(5), 276–283, (2003).

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

23. S. Sakamoto, S. Hongo, T. Okamoto, Y. Iwaya, and Y. Suzuki, "Sound-space recording and binaural presentation system based on a 252-channel microphone array," *Acoust. Sci. & Tech.*, 36(6), 516–526, (2015).