MADRID
inter.noise 2019
June 16 - 19
NOISE CONTROL FOR A BETTER ENVIRONMENT

Binaural synthesis using a spherical microphone array based on the solution to an inverse problem

Shuichi, Sakamoto[1], César Salvador[2], Jorge Treviño[3], and Yôiti Suzuki[4]
Research Institute of Electrical Communication and Graduate School of Information Sciences, Tohoku University
2-1-1 Katahira, Aoba-ku, Sendai, Miyagi 980-8577, Japan

ABSTRACT

Binaural technologies can convey rich spatial auditory information to listeners, using simple equipment such as headphones. Head-related transfer function (HRTF) datasets and spherical microphone arrays are important components to realize advanced binaural recording and reproduction methods. One of these methods is binaural Ambisonics, which captures and reproduces a three-dimensional sound space based on a spherical harmonic analysis. We have proposed an advanced binaural reconstruction method without relying on an intermediate sound field representation. In this method, known as SENZI (Symmetrical object with ENchased ZIllion microphones), individualized binaural signals are directly generated from the microphone recordings by applying weighting filters, which are calculated by inverting a linear system. The accuracy of the sound space synthesized by SENZI is higher at the head-shadow region than that synthesized by conventional binaural Ambisonics. However, the number of microphones imposes a limit on the accuracy at higher frequencies according to sampling theory. This limit can be overcome by making assumptions on the recorded contents or using perceptual information. Because the sensitivity to the phase of human spatial hearing decreases at higher frequencies, we propose utilizing this perceptual knowledge to reconstruct high-frequency contents. Simulation results revealed that high-frequency amplitudes are accurately reproduced using the proposed approach.

Keywords: Spherical acoustics, Binaural technology, Spherical microphone array, Ambisonics
I-INCE Classification of Subject Number: 72

1. INTRODUCTION

Advanced three-dimensional (3D) sound space recording and reproduction methods are crucial for realizing high-definition multimedia systems. One promising method in this direction is binaural technology. Binaural technologies can be realized using simple equipment, such as headphones. An important component of such

[1] saka@ais.riec.tohoku.ac.jp
[2] salvador@ais.riec.tohoku.ac.jp
[3] jorge@ais.riec.tohoku.ac.jp
[4] yoh@riec.tohoku.ac.jp

technologies is the head-related transfer function (HRTF). HRTFs comprehensively contain individual auditory cues, which are important to present sounds at accurate locations outside of the head [1]. The acquisition subsystem is also a highly important component for realizing high-definition binaural technologies. Spherical microphone arrays with numerous microphones are typically employed to design binaural systems with these two components, because they enable us to analyze the sound space easily using spherical harmonic decomposition. This class of technologies can be classified into two general types: HRTF modeling and microphone signal modeling approaches [2].

We have been developing signal processing methods that generate individualized binaural signals directly from microphone recordings, without relying on intermediate sound field representations, such as by spherical harmonic decomposition. This approach, known as SENZI (Symmetrical object with ENchased ZIllion microphones) [3,4], applies a set of weighting coefficients to recorded microphone signals, resulting in a target spatial pattern defined by the HRTF dataset for a specific listener. The weighting coefficients are calculated by inverting a linear system between transfer functions of the rigid sphere and HRTF dataset. To obtain accurate sound space information over the whole audible frequency range, numerous microphones have to be set on the rigid sphere. However, it is cost-consuming to develop spherical microphone arrays with densely distributed microphones. The number of microphones imposes a limit on the accuracy at higher frequencies, according to spatial sampling theory. Therefore, it is important to synthesize perceptually accurate sound space information by using a spherical microphone array with fewer microphones on the sphere.

To overcome this limit, knowledge on a human’s sound space perception could be effectively exploited to develop an advanced SENZI method. Various attempts have been proposed to enhance the perceptual accuracy of a synthesized sound space [5,6] for the same number of microphones set on a sphere. We focus on the insensitivity of human phase perception at high frequency components. In this study, the accuracy of the amplitude responses of HRTFs was investigated by modifying the phase information of HRTFs at high frequency regions, where the effect of spatial aliasing is observed.

2. SENZI: SOUND SPACE SYNTHESIS USING A SPHERICAL MICROPHONE ARRAY [3,4]

Our developed method, SENZI, represents an HRTF modeling approach [2]. To calculate and synthesize a listener's HRTF using inputs from multiple spatially distributed microphones, recorded signals from each microphone are simply weighted and summed to synthesize a listener's HRTF. The weights can be changed according to a listener's 3D head movements. Owing to this feature, 3D sound-space information is accurately acquired, irrespective of head movements. 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.

A spherical microphone array using the actual SENZI system is placed at a radius of 8.5 cm. Furthermore, 252 microphones are installed on this spherical object. The position of each microphone is calculated on the basis of a regular icosahedron. The intervals between all neighboring microphones are almost the same: the average interval is 2.04 cm. Consequently, the limit for the array's spatial resolution appears at a frequency of over approximately8.5 kHz. A small digital omnidirectional electric condenser microphone (ECM) (KUS5147; Hosiden Co., Ltd.) is set at one of the 252 calculated positions.

3. PREPROCESSING OF HRTF PHASE AT HIGH FREQUENCY REGIONS [6]

As mentioned in the introduction, the accuracy of the synthesized auditory space information depends strongly on the interval of microphones on the rigid sphere. When a spherical microphone array at the radius of 8.5 cm is employed to record sound space information, the interval of each microphone should be set less than 1 cm to cover the whole audible frequency range without the distortion due to the effect of spatial-aliasing. This means that approximately 1,000 microphones are required to record the sound space. We developed a 252-ch spherical microphone array at a radius of 8.5 cm. The interval of each microphone is approximately 2 cm. Therefore, the accuracy of the sound space synthesized using this microphone array is degraded at a frequency above approximately 8.5 kHz owing to spatial aliasing.

To relax this restriction, knowledge of human auditory spatial perception is incorporated. The interaural phase difference (IPD) and interaural level difference (ILD) can be utilized as powerful cues to perceive the horizontal sound space. However, the IPD cue becomes less relevant as the frequency increases. Therefore, by modifying the phase information at high-frequency regions, we attempt to increase the perceptual accuracy of the synthesized sound space. The amplitude responses of HRTFs were accurately synthesized at high-frequency regions.

In this study, the time alignment with the original HRTFs [6] was applied to the HRTFs used to calculate the weighting coefficients of the SENZI method. An arbitrary set of HRTFs defined at the discrete sound source directions $\Omega_p \equiv (\phi_p, \theta_p) \in \boldsymbol{S}^2$, with $p = \{1, \ldots, P\}$ as the index of the available grids points, is expressed as follows:

$$\boldsymbol{h}_p(\omega) = \left[h^l(\Omega_p, \omega), h^r(\Omega_p, \omega)\right]^T.$$

Here, the time-aligned HRTF set is computed as

$$h\ l, r(\Omega_p, \omega) = h^{l,r}(\Omega_p, \omega) A_p^{l,r}(\omega),$$

where the frequency response of the all-pass filter $A_p^{l,r}(\omega)$ is defined as

$$A_p^{l,r}(\omega) = \begin{cases} 1 & \text{for } \omega < \omega_c \\ e^{-i(\omega-\omega_c)\tau_p^{l,r}} & \text{for } \omega \geq \omega_c, \end{cases}$$

where $\omega_c = 2\pi f_c$, $i = \sqrt{-1}$. The time offset $\tau_p^{l,r}$is calculated as follows:

$$\tau_p^r = \cos(\theta_p)\sin(\phi_p)\, r_H c^{-1}, \tau_p^l = -\tau_p^r,$$

where $c = 343$ (m/s) is the speed of sound and $r_H$is a head radius of 8.5 cm. By applying these processes to the original HRTFs, the time difference between the center of the head and the ears can be compensated for each grid point p.

4. EVALUATION

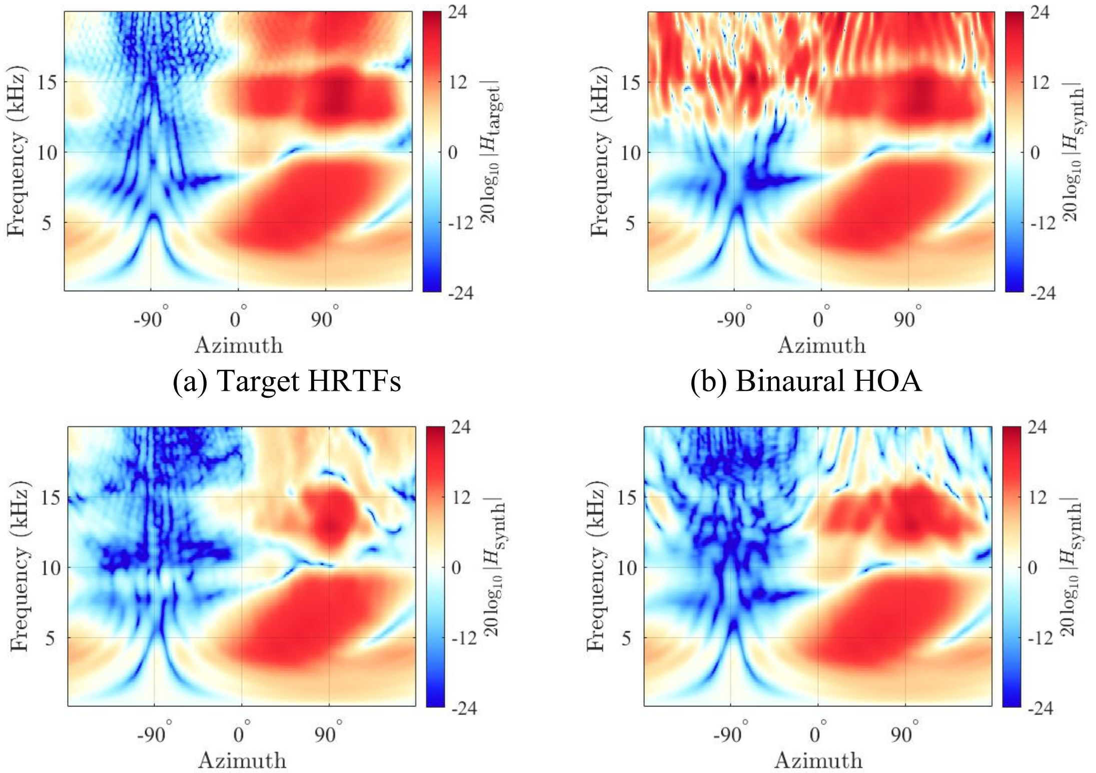
To evaluate the effect of time-alignment HRTFs on the accuracy of a synthesized sound space, the obtained time-alignment HRTFs were applied to the SENZI method to calculate the weighting coefficients, and HRTFs on the horizontal plane were synthesized.

The HRTFs of a dummy-head (SAMRAI; Koken Co. Ltd.) were utilized as the target to be realized to evaluate the performance. The accuracy of the synthesized sound-space was analyzed in terms of the synthesized HRTFs for a point sound source using a computer simulation. In the simulation, 2,562 HRTFs were used to calculate the weighting coefficients of the SENZI method. The 2,562 sound source positions were determined as follows. First, a regular icosahedron inscribed in a sphere of radius 1.5 m was assumed. Next, each surface of the regular icosahedron was divided into 256 small equilateral triangles. The apices of these triangles were projected to the surface of the sphere. The results show that 2,562 apices were obtained at a distance of 1.5 m from the center of the sphere, and these were employed as sound source positions. HRTFs from the obtained sound positions were calculated numerically using the boundary element method [7]. For the spherical microphone array, 252-ch microphones were set on a rigid sphere at a radius of 8.5 cm. The specification of the spherical microphone array is the same as that previously developed as an actual microphone array [4]. Because the sensitivity of the phase information is degraded when the frequency is above approximately 1.5 kHz, time-alignment was applied to the target HRTFs at frequencies over 1.5 kHz ($\omega_c = 1.5$ kHz). To compare the accuracy of the sound space information synthesized by high-order Ambisonics (HOA), the calculated time-alignment HRTFs were also applied to binaural Ambisonics [6].

Figure 1 depicts the synthesized HRTFs using both the original and time-alignment HRTFs for calculating the weighting coefficients of the SENZI method. At around the middle-frequency range (10–15 kHz), some peaks appear to be accurately simulated by using time-alignment HRTFs with the SENZI method. On the other hand, at frequencies over 15 kHz the accuracy of the synthesized HRTFs is degraded compared with the conventional SENZI method and binaural HOA.

Figure 2 illustrates the reproduction error of the synthesized HRTFs using the SENZI method. As shown in Fig. 1, the reproduction error at the middle-frequency range is clearly decreased, while a large error is observed in the high-frequency region. Moreover, the frequency response in the head-shadow region can be accurately synthesized using the SENZI method compared with binaural HOA.

Although it is unclear why the accuracy of the synthesized HRTFs is degraded in the high-frequency region by applying the time-alignment HRTFs, this is not a serious problem. Because the weighting coefficients are calculated independently at each frequency, the accuracy can easily be recovered using appropriate HRTFs. Indeed, by



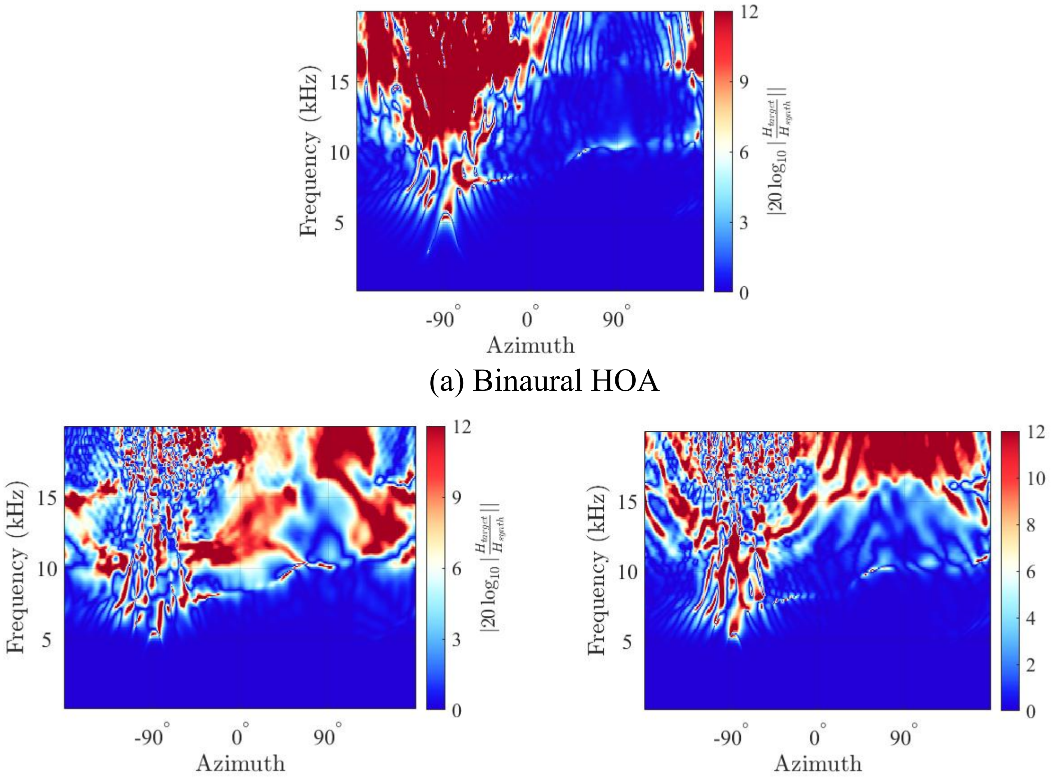
(a) Target HRTFs

(b) Binaural HOA

(c) Conventional SENZI method

(d) Revised SENZI method

Fig. 1 Azimuthal patterns of synthesized HRTFs using SENZI method and binaural HOA.



(a) Binaural HOA

(b) Conventional SENZI method

(c) Revised SENZI method

Fig. 2 Reproduction error of synthesized HRTFs using the SENZI method and binaural HOA

using the original HRTFs over 15 kHz, a high accuracy can be maintained in the high frequency region. It will be important in future work to investigate how to select the optimum HRTFs to increase the accuracy of the amplitude responses of synthesized HRTFs in high-frequency regions.

5. CONCLUSION

Binaural recording and reproduction methods using head-related transfer functions and spherical microphone arrays can convey rich spatial auditory information to listeners. Our proposed SENZI method represents an HRTF modeling approach, and generates individual binaural signals by applying weighting filters, which are calculated by inverting a linear system. However, the number of microphones imposes a limit on the accuracy at higher frequencies, according to sampling theory.

By introducing knowledge on human spatial perception, especially the insensitivity of phase perception in high-frequency regions, the accuracy of the amplitude responses of HRTFs synthesized using the SENZI method was improved in the middle-frequency region. Moreover, HRTFs synthesized using the SENZI method were more accurate than those synthesized by binaural HOA in the head-shadow region.

It would be possible to improve the amplitude responses of synthesized HRTFs in high frequency regions more accurately by adding appropriate phase information. ILD also provides a good index for improving the perceptual accuracy of sound space information in high-frequency regions. We plan to introduce such information into the SENZI method, and improve the perceptual accuracy of sound spaces generated by the SENZI method.

6. ACKNOWLEDGEMENTS

A part of this work was supported by JSPS KAKENHI Grant Numbers JP16H01736 and the Cabinet Office, Government of Japan, Cross-ministerial Strategic Innovation Promotion Program (SIP). We also wish to thank Mr. Ryo Togashi for his help with the analysis of the results.

7. REFERENCES

1. Morimoto M. and Ando Y., *"On the simulation of sound localization"*, J. Acoust. Soc. Jpn. (E), 1(3), 167-174 (1980)
2. Salvador C. D., Sakamoto S., Treviño J., and Suzuki Y. *"Design theory for binaural synthesis: combining microphone array recordings and head-related transfer function datasets"*, Acoust. Sci. & Tech., 35(3), 159-165 (2014)
3. Sakamoto S., Hongo S., and Suzuki Y., *"3D sound-space sensing method based on numerous symmetrically arranged microphones"*, IEICE Trans. Fundamentals, E97-A(9), 1893-1901 (2014)
4. Leo L. Beranek and István L. Vér, *"Noise and Vibration Control Engineering – Principles and Applications"*, edited by Leo L. Beranek and István L. Vér, John Wiley & Sons, New York (2006)
5. Rasumow E., Hansen M., van de Par S., Püschel D., Mellert V., Doclo S., and Blau M., *"Regularization approaches for synthesizing HRTF directivity patterns"*, IEEE/ACM Trans. Audio, Speech, Language Process., 24(2), 215-225 (2016)
6. Zaunschirm M., Schörkhuber C., and Höldrich R., *"Binaural rendering of Ambisonic signals by head-related impulse response time alignment and a diffuseness constraint"*, J. Acoust. Soc. Am., 143, 3616-3627 (2018)
7. Otani M., and Ise S., *"Fast calculation system specialized for head-related transfer function based on boundary element method"*, J. Acoust. Soc. Am., 119, 2589-2598 (2006)