社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

Enhancing the binaural synthesis from spherical microphone array recordings by using virtual microphones

César SALVADOR†, Shuichi SAKAMOTO†, Jorge TREVIÑO†, and Yôiti SUZUKI†

† Research Institute of Electrical Communication and Graduate School of Information Sciences, Tohoku University, Sendai, Japan

E-mail: {salvador,saka,jorge}@ais.riec.tohoku.ac.jp, yoh@riec.tohoku.ac.jp

Abstract Binaural synthesis from rigid spherical microphone arrays requires high spatial resolutions. Physically adding microphones to available arrays, however, is not always feasible. In environments such as conference rooms or concert halls, prior knowledge about the source positions allows predicting microphone signals by relying upon a physical model for the acoustically rigid sphere. Recently, we have used this model to formulate and evaluate a method that enhances spatial sound recordings by adding virtual microphones to the array. In this study, we apply this method to enhance the spatial accuracy of binaural synthesis when it is performed in anechoic and reverberant conditions.

Key words 3D audio technology, binaural synthesis, head-related transfer functions, spherical microphone arrays, virtual microphones.

1. Introduction

Spatial audio technologies for the recording of acoustic pressure signals at the eardrums [1], namely binaural signals, are becoming highly demanded due to the popularity of spatial audio reproduction devices for personal use, such as headphones [2] and personal sound zone systems [3]. When aiming to capture an acoustic environment for its reconstruction in a remote place, an important characteristic of a binaural recording system is its spatial accuracy or ability to resolve sounds along directions. Spatial accuracy indeed has a high impact on the degree of realism and naturalness when aiming to re-create the experience of being immersed in a distinct acoustic environment [4].

A modern type of binaural systems [5–12] aims to synthesize the binaural signals by combining the spatial information available in the recordings made with a spherical microphone array mounted on a rigid baffle [13–15] and the spatial information available in acoustical characterizations of the external anatomical shapes of individual listeners. These acoustical human models are characterized for a set of surrounding positions and constitute datasets of the so-called head-related tranfer functions (HRTFs) [16]. The HRTFs are linear filters describing the transmission of sound from a position in space to the listener's eardrums.

The use of rigid spherical microphone arrays allows to consider the dynamic auditory cues that correspond to the head movements [17]. The use of HRTF datasets, on the other hand, allows to consider the individual auditory cues that arises from the interactions of sound waves with the external anatomical shapes of the listeners [16]. Obtaining HRTFs for dense sets of positions is possible thanks to recent techniques based on 3D model acquisition systems and numerical acoustics methods such as the boundary element method (BEM) [18]. The recording of high-definition spatial sound to ensure the accurate synthesis of binaural cues, however, still represents a challenge because it demands spherical arrays with a large number of microphones [4, 9–12].

The costs involved in the construction of high spatial resolution microphone arrays have confined their use to research purposes [6–8], and the maximum spatial resolutions of commercially available arrays are still in the range of a few tens of microphones [13, 14]. Moreover, physically increasing the spatial resolution of available arrays is not always feasible.

In environments such as conference rooms or concert halls, however, the positions of sound sources are often conveniently confined to a small region of space. In these particular conditions, when the sound source positions can be assumed to be known, the pressure generated at any point on the rigid spherical baffle where the microphones are mounted can be estimated using a physical model of the rigid sphere [19, 20].

Recently, we have proposed a virtual microphone generation method to enhance the spatial resolution of rigid spherical microphone array recordings [21]. The method takes advantage of prior knowledge about the source positions to generate recording signals at positions on the baffle where there are no microphones, and this is done by relying on the acoustically rigid sphere model.

In this study, we apply the virtual microphone generation

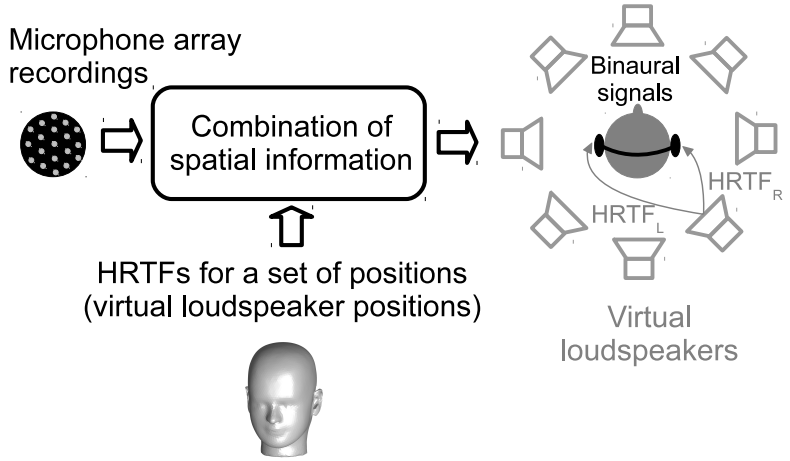

Fig. 1 Binaural synthesis from microphone array recordings and HRTF datasets. The use of microphone arrays allow for the consideration of dynamic auditory cues and multiple moving listeners. The use of HRTF datasets allow for the consideration of individual auditory cues.

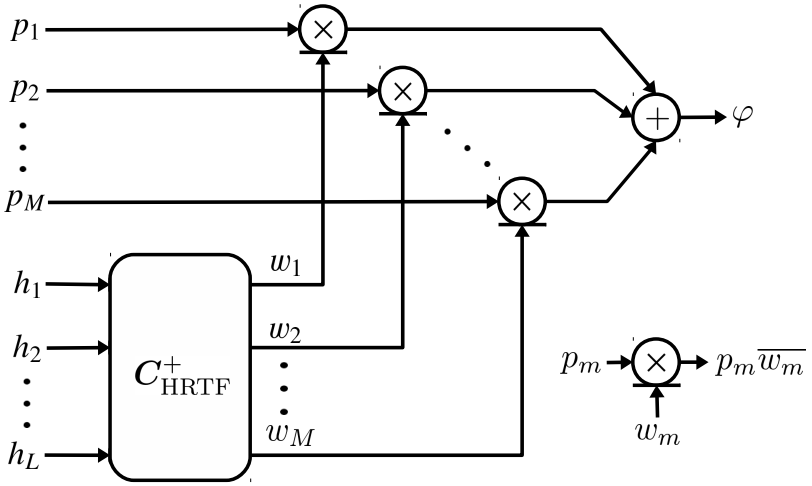

Fig. 2 Combination of microphone array recordings and HRTF datasets based on the HRTF spatial pattern modeling approach [7, 8, 12]. The combination matrix $\boldsymbol{C}_{\mathrm{HRTF}}$ contains all possible acoustic transfer functions between virtual loudspeaker positions and microphone positions. Matrix $\boldsymbol{C}_{\mathrm{HRTF}}$ needs to be inverted to generate $\boldsymbol{C}_{\mathrm{HRTF}}$.

method to enhance the spatial accuracy of modern binaural systems by adding virtual microphones to the rigid spherical microphone array. In the remainder of this paper we present an overview of modern binaural systems and then we describe the virtual microphone generation method. We also present numerical examples when enhanced binaural synthesis is performed in anechoic and reverberant conditions.

2. Binaural synthesis from microphone array recordings and HRTF datasets

During the last decade, several binaural synthesis methods that combine microphone array recordings and HRTF datasets have been studied [5–12]. They aim at binaurally rendering a sound pressure field sampled at the positions for which a set of HRTFs was obtained. In other words, these methods aim to synthesize the binaural signals due to an array of virtual loudspeakers placed at the positions used to obtain the HRTFs (see Fig. 1).

One approach for binaural synthesis can be regarded as a beamformer [7, 8, 12]. An individual HRTF dataset ($\boldsymbol{h}$) constitutes a specified spatial pattern to be approximated by a set of weighting filters ($\boldsymbol{w}$) that will be applied to the microphone recordings ($\boldsymbol{p}$). The weighting filters are calculated by solving a linear system of equations that approximate the HRTF dataset. The entries of the matrix associated ($\boldsymbol{C}_{\mathrm{HRTF}}$) to the linear system are acoustic transfer functions from the positions of microphones to the positions used to obtain the HRTF dataset.

The linear system to be solved is formulated by linearly combining the acoustic transfer functions in $\boldsymbol{C}_{\mathrm{HRTF}}$ so as to approximate the HRTF dataset $\boldsymbol{h}$:

$$\boldsymbol{C}_{\mathrm{HRTF}}\boldsymbol{w} = \boldsymbol{h} + \epsilon_{\mathrm{HRTF}}. \tag{1}$$

Here, ϵ_{HRTF} denotes the approximation error.

The synthesis algorithm obtained from the solution to (1) can be summarized in two steps:

（1） Weighting filters $\boldsymbol{w} = \boldsymbol{C}^{+}_{\text{HRTF}}\boldsymbol{h}$.

（2） Binaural signals $\varphi = \overline{\boldsymbol{w}^{\top}}\boldsymbol{p}$.

The matrix organization used here is detailed below.

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.

The recordings of an array composed of M microphones are organized in the vector

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

Each entry p_m of $\boldsymbol{p}$, where $m = 1, 2, ..., M$, represents a sample of sound pressure recorded at a microphone position $\vec{a}_m$ on the spherical rigid baffle.

Finally, the HRTFs of the dataset are organized in the matrix

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

Each entry h_ℓ^{left} or h_ℓ^{right} of $\boldsymbol{h}$, where $\ell = 1, 2, ..., L$, represents a sample of the free-field HRTF for the left or right ear, respectively. Each entry of $\boldsymbol{h}$ is characterized for a virtual loudspeaker position $\vec{b}_\ell$.

The algorithm above described results in the array processing structure shown in Fig. 2.

3. Virtual microphone generation method for spatial resolution enhancement

This section overviews our recently proposed method [21] to increase the total number M of microphone array signals by adding virtual microphones to the array. The method uses prior knowledge about the source positions to generate recording signals at positions without microphones. This is done by relying on the acoustically rigid sphere model.

Consider a rigid sphere of radius a and a sound source at a distance $r > a$ measured from the center of the rigid sphere. The total pressure generated by a sound source placed at a position $\vec{r}$ and measured by an ideal microphone placed at a position $\vec{a}$ on the surface of the rigid sphere is defined [19]:

$$P(\vec{r}, \vec{a}, k) = \frac{-1}{ka^2} \sum_{n=0}^{\infty} \frac{h_n(kr)}{h_n'(ka)} (2n+1) L_n(\cos\Theta_{\vec{r},\vec{a}}), \tag{5}$$

where

$$k = \frac{2\pi f}{c}. \tag{6}$$

Here, f denotes the frequency and c denotes the speed of sound in air. In (5), h_n denotes the spherical Hankel function of order n and the symbol $'$ indicates derivative with respect to the argument. In addition, L_n denotes the Legendre polynomial of order n evaluated at the cosine of the angle $\Theta_{\vec{r},\vec{a}}$ between $\vec{r}$ and $\vec{a}$.

The model in (5) is used to relate the pressure at two arbitrary points $\vec{a}_m$ and $\vec{a}_v$ on the rigid sphere when a reference sound source position $\vec{r}_{\text{ref}}$ is assumed. This defines the following surface pressure variation function:

$$F_{m\mapsto v} = \frac{P_v}{P_m}, \tag{7}$$

where

$$P_m = P(\vec{r}_{\text{ref}}, \vec{a}_m, k), \text{ and } P_v = P(\vec{r}_{\text{ref}}, \vec{a}_v, k). \tag{8}$$

The function $F_{m\mapsto v}$ represents the transmission of sound on the surface of the rigid sphere from one arbitrary point $\vec{a}_m$ to another arbitrary point $\vec{a}_v$. This function can also be regarded as a surface pressure interpolation filter and used to define a virtual microphone generation method. This function is used for the generation of virtual microphone signals in the vicinity of real microphones in the array. The vicinity is considered around a single real microphone in the method below described.

3.1 Surface pressure interpolation from a single nearest microphone

Let p_m denote the signal recorded by a real microphone placed at a position $\vec{a}_m$ in the array. Let p_v denote a virtual microphone signal to be generated at a position $\vec{a}_v$ in the array where there are no microphones. The virtual microphone signal p_v is generated by applying the filter $F_{m\mapsto v}$ to p_m as follows:

$$p_v = F_{m\mapsto v} \times p_m. \tag{9}$$

Note that the application of (9) requires the specification of two positions to calculate $F_{m\mapsto v}$ in accordance to (7). These are the reference position $\vec{r}_{\text{ref}}$ and the virtual microphone position $\vec{a}_v$. The position $\vec{a}_m$ of the real microphone from which interpolation is performed can be simply selected by a nearest neighbor search. This search consists of two steps. First, the angles between $\vec{a}_v$ and each real microphone in the array are calculated. Next, the microphone in the array that creates the smallest angle from $\vec{a}_m$ is selected.

4. Enhancement of binaural synthesis by adding virtual microphones

In this section we present numerical examples of virtual microphone generation applied to the binaural synthesis scheme shown in Fig. 2. The method is used to increase the number of microphone signals. Given the rotational symmetry of the rigid sphere, the results presented hereafter consider the horizontal plane as a representative case for the whole sphere. All examples uses a number $L = 30$ of HRTFs for virtual loudspeakers at a distance 150 on the horizontal plane. The rigid spherical baffle where the microphones are mounted is of radius 8.5 cm. Two examples

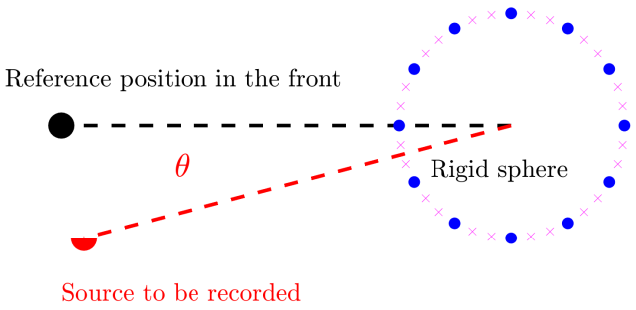

Fig. 3 Geometry for virtual microphone generation. Blue dots indicate real microphone positions. Magenta marks indicate virtual microphone positions. The angle θ is the azimuth on the horizontal plane.

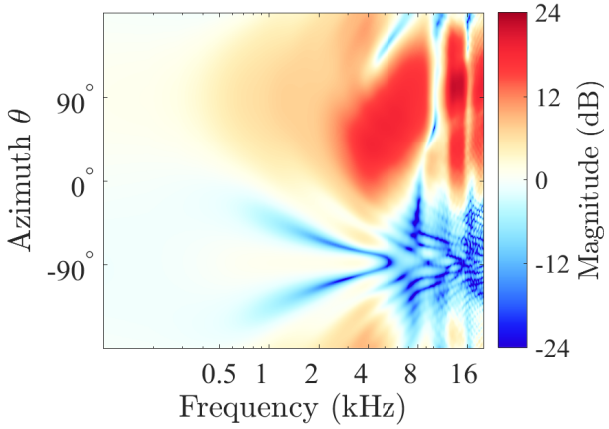

Fig. 4 Target signals for sound sources on the horizontal plane.

of number of real microphones are considered, $M = 30$ and $M = 120$. The microphones are mounted on a horizontal circle. The reference position $\vec{r}_{\text{ref}}$ to calculate the surface pressure variation filters $F_{m\mapsto v}$ was set in the front of the listener ($\theta = 0°$) at a 150 cm distance (see Fig. 3).

Figure 4 presents left-ear HRTFs for sources on the horizontal plane. In the numerical experiments, these signals constitute the target binaural signals for the left ear. Figures from 5 to 7 show the synthesized binaural signals when recordings are modeled in anechoic conditions, whereas Figs. from 8 to 10 present synthesis examples for recordings made in reverberant conditions.

Figures 5 and 6 show synthesis examples using $M = 30$ and $M = 120$ real microphones, respectively. It is observed that increasing the number of real microphones improves the performance at higher frequencies.

Figure 7 shows the synthesis results for $M = 30$ real microphones and 90 additional virtual microphones. When contrasting these results with Figs. 5 and 6, it is observed that adding virtual microphones also improves the performance at higher frequencies in anechoic conditions. However, this is obtained at the cost of degrading the accuracy for sources on the opposite side of the reference position, that is, in positions behind the listener.

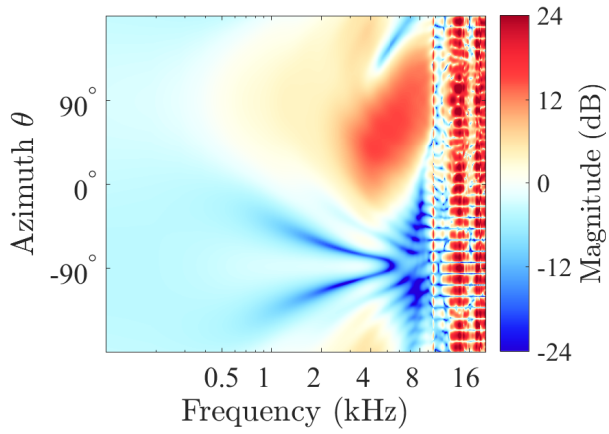

Fig. 5 Synthesis in anechoic conditions using $L = 30$ HRTFs and $M = 30$ real microphones.

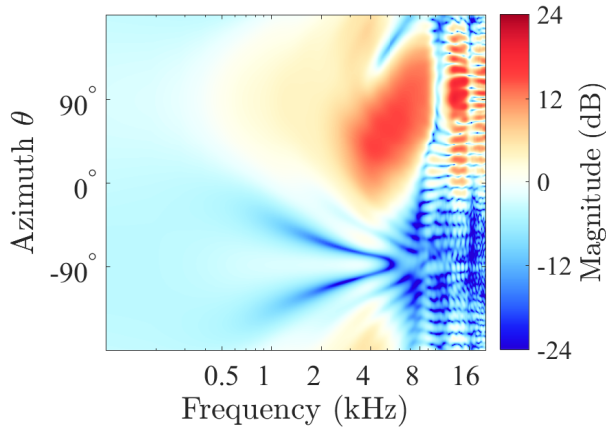

Fig. 6 Synthesis in anechoic conditions using $L = 30$ HRTFs and $M = 120$ real microphones.

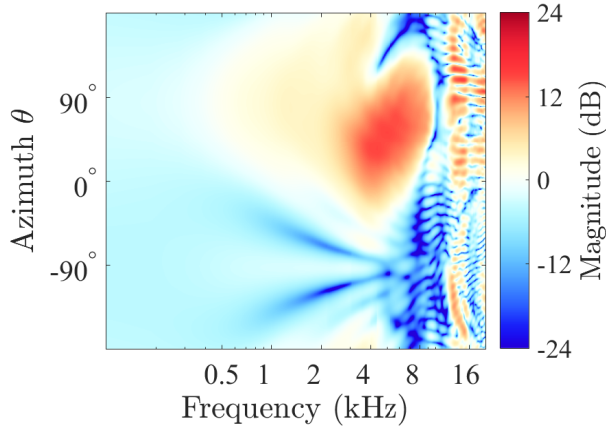

Fig. 7 Synthesis in anechoic conditions using $L = 30$ HRTFs, $M = 30$ real microphones, and 90 additional virtual microphones (120 microphones in total).

Recording in a reverberant condition was simulated using the algorithm in [20] to describe a rigid sphere placed inside a rectangular parallelepiped room. The image method was used to describe a high-order reflection model of the room. The center of

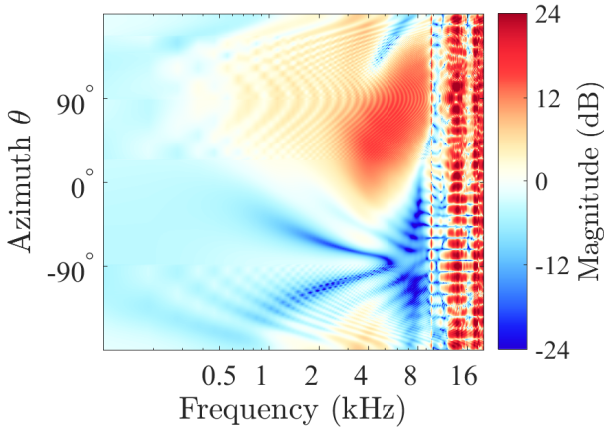

Fig. 8 Synthesis in reverberant conditions using $L = 30$ HRTFs, $M = 30$ real microphones.

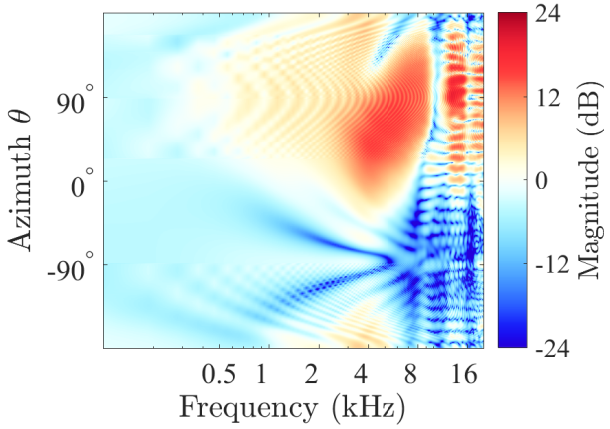

Fig. 9 Synthesis in reverberant conditions using $L = 30$ HRTFs, $M = 120$ real microphones.

the coordinates coincides with a bottom corner of the room. The dimensions of the room included a 5 m width (along x), 6 m length (along y), and 4 m height (along z). The reflection coefficients of all walls were 0.3. The center position of the microphone array was $(1.6, 4.05, 1.7)$ m. The front position lies along the positive x-axis.

Figures 8 and 9 show synthesis examples when recordings were made in reverberant conditions by using $M = 30$ and $M = 120$ real microphones, respectively. Similarly to the results obtained in anechoic conditions, it is observed that increasing the number of real microphones also extends the range of operation towards the higher frequencies.

Finally, Fig. 10 shows the synthesis results in reverberant conditions when $M = 30$ real microphones were used and 90 virtual microphones were added. When contrasting these results with the results shown in Figs. 8 and 9, it is observed that adding virtual microphones improves the performance at higher frequencies. However, similarly to the anechoic case, a higher frequency range of operation is achieved at the cost of degrading the accuracy for sources behind the listener, which is the region lying

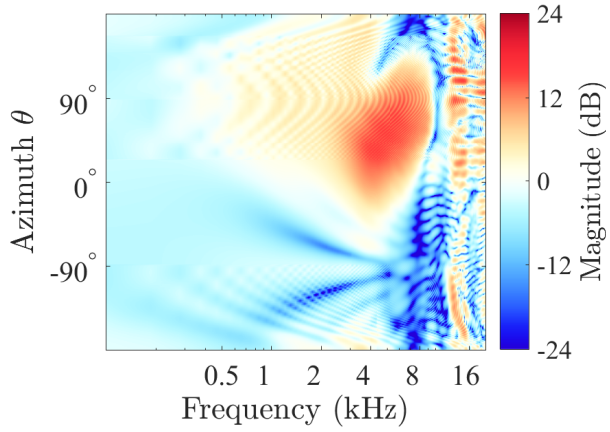

Fig. 10 Synthesis in reverberant conditions using $L = 30$ HRTFs, $M = 30$ real microphones, and 90 additional virtual microphones (120 microphones in total).

on the opposite side of the reference position.

5. Conclusion

We presented a method to generate virtual microphone signals for the enhancement of the spatial resolution of rigid spherical microphone arrays. Our method requires prior knowledge about source positions. It can be used in recording environments where the source positions are often confined to a small region of space, such as conference rooms or concert halls.

The method was applied to binaural synthesis with circular microphone arrays. Evaluations were presented by means of simulations. The results show that adding virtual microphones has the benefit of extending the frequency range of operation towards the higher frequencies. However, this is obtained at the cost of decreasing the spatial accuracy in the opposite side to the assumed source positions.

Extensions of this work can include improvements in phase unwrapping algorithms, surface pressure interpolation from multiple nearest microphones, and further considerations of surface pressure interpolation under diffuse fields.

6. Acknowledgments

Part of this study was supported by the JSPS Grant-in-Aid for Scientific Research no. JP16H01736 and no. JP17K12708.

References

[1] H. Moller, “Fundamentals of binaural technology,” *Appl. Acoust.*, vol. 36, no. 3-4, pp. 171–218, 1992.

[2] V. Algazi and R. Duda, “Headphone-Based Spatial Sound,” *IEEE Signal Process. Mag.*, vol. 28, no. 1, pp. 33–42, Jan. 2011.

[3] T. Betlehem, W. Zhang, M. Poletti, and T. Abhayapala, “Personal sound zones: Delivering interface-free audio to multiple listeners,” *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 81–91, Mar. 2015.

[4] 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.*, vol. 133, no. 5, pp. 2711–2721, May 2013.

[5] V. R. Algazi, R. O. Duda, and D. M. Thompson, “Motion-tracked

binaural sound," *J. Audio Eng. Soc.*, vol. 52, no. 11, pp. 1142–1156, 2004.

[6] 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," in *AES 119*, New York, USA, Oct. 2005.

[7] 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, pp. 1893–1901, Sep. 2014.

[8] 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. Technol.*, vol. 36, no. 6, pp. 516–526, 2015.

[9] C. D. Salvador, S. Sakamoto, J. Trevino, J. Li, Y. Yan, and Y. Suzuki, "Accuracy of head-related transfer functions synthesized with spherical microphone arrays," *Proc. Mtgs. Acoust.*, vol. 19, no. 1, Apr. 2013.

[10] C. D. Salvador, S. Sakamoto, J. Trevino, and Y. Suzuki, "Numerical evaluation of binaural synthesis from rigid spherical microphone array recordings," in *Proc. AES Int. Conf. Headphone Technology*. Aalborg, Denmark: Audio Engineering Society, Aug. 2016.

[11] ——, "Spatial accuracy of binaural synthesis from rigid spherical microphone array recordings," *Acoust. Sci. Technol.*, vol. 38, no. 1, pp. 23–30, Jan. 2017.

[12] ——, "Design theory for binaural synthesis: Combining microphone array recordings and head-related transfer function datasets," *Acoust. Sci. Technol.*, vol. 38, no. 2, pp. 51–62, Mar. 2017.

[13] J. Meyer and G. Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield," in *Proc. IEEE ICASSP*, vol. II, Orlando, FL, USA, May 2002, pp. 1781–1784.

[14] B. Rafaely, "Analysis and design of spherical microphone arrays," *IEEE Trans. Speech, Audio Process.*, vol. 13, no. 1, pp. 135–143, Jan. 2005.

[15] C. D. Salvador, S. Sakamoto, J. Trevino, and Y. Suzuki, "Boundary matching filters for spherical microphone and loudspeaker arrays," *IEEE/ACM Trans. Audio, Speech, Language Process.*, 2017, accepted with minor revisions.

[16] J. Blauert, *Spatial hearing: The psychophysics of human sound localization*, revised ed. Cambridge, MA, USA; London, England.: MIT Press, 1997.

[17] Y. Iwaya, Y. Suzuki, and D. Kimura, "Effects of head movement on front-back error in sound localization," *Acoust. Sci. Technol.*, vol. 24, no. 5, pp. 322–324, 2003.

[18] M. Otani and S. Ise, "Fast calculation system specialized for head-related transfer function based on boundary element method," *J. Acoust. Soc. Am.*, vol. 119, no. 5, pp. 2589–2598, May 2006.

[19] R. O. Duda and W. L. Martens, "Range dependence of the response of a spherical head model," *J. Acoust. Soc. Am.*, vol. 104, no. 5, pp. 3048–3058, Nov. 1998.

[20] D. P. Jarret, E. A. P. Habets, M. R. P. Thomas, and P. A. Naylor, "Rigid sphere room impulse response simulation: algorithm and applications," *J. Acoust. Soc. Am.*, vol. 132, no. 3, pp. 1462–1472, Sep. 2012.

[21] C. D. Salvador, S. Sakamoto, J. Trevino, and Y. Suzuki, "Enhancement of spatial sound recordings by adding virtual microphones to spherical microphone arrays," *J. Inf. Hiding and Multimedia. Signal Process.*, no. Special Issue on Enrichment of sound, speech and music media, 2017, in print.