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.

A model for spatial sound systems comprising sound field recording, spatial editing, and binaural reproduction

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 We present a mathematical model to combine and edit spatial sound information defined on the sphere. Such model aim to enable the development of binaural systems capable of recording far sounds in all directions. Recorded far sounds can subsequently be discriminated along directions, and independently brought closer to the listener. During binaural reproduction, sounds can be heard as if they were closer than their original distance. Illustrative examples based on spherical arrays are provided by means of simulations.

Key words 3D audio technology, binaural systems, head-related transfer functions, microphone arrays, transform-domain acoustics.

1. Introduction

Spatial sound systems are a key component in the development of realistic multimedia telecommunication technologies that aim to convey the perceptual experience of being immersed in a distinct environment. The spatial features of sound are indeed essential to enhance the levels of perceived presence and naturalness in the technology-mediated experience [1, 2].

Binaural systems [3, 4] are a class of spatial sound systems that aim to evoke the experience of being immersed in an acoustic environment by synthesizing the sound pressure at the eardrums of listeners, namely binaural signals. The synthesized binaural signals can subsequently be presented through binaural devices such as headphones [5] and personal sound zone systems [6].

A type of recent binaural systems [7–13] aims to synthesize the binaural signals by combining the spatial information available in microphone array recordings and individual head-related transfer function (HRTF) datasets. An individual HRTF dataset is composed of linear filters describing the transmission of sound from a set of positions in space to the eardrums of an individual listener [14].

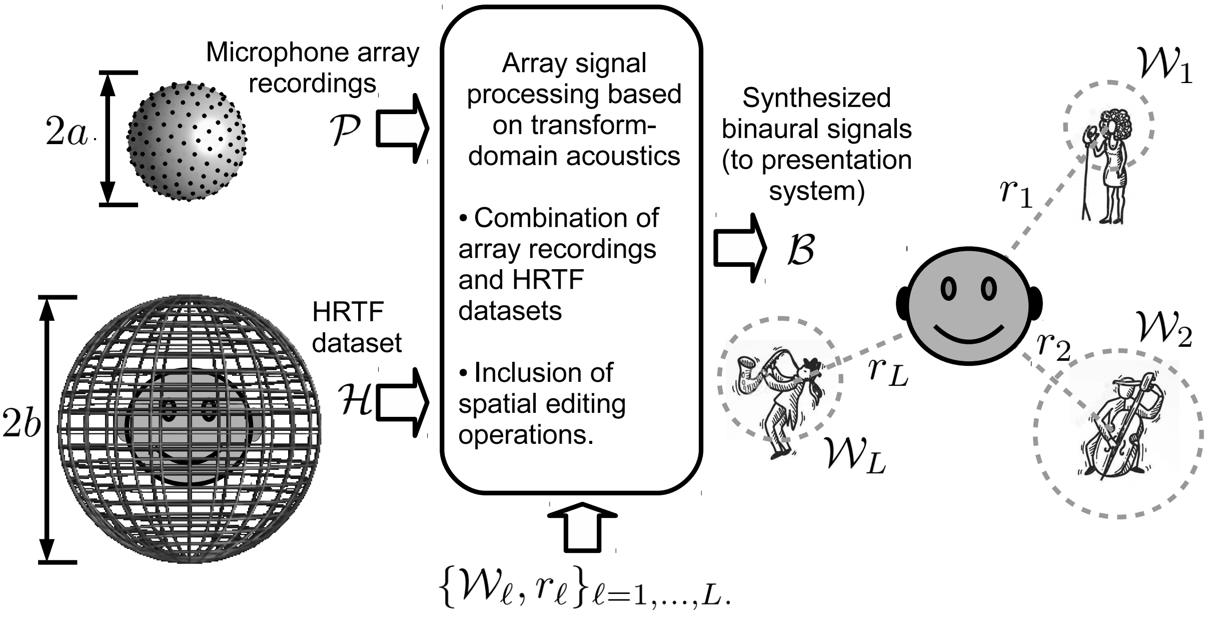
Combination of array recordings and HRTF datasets is performed relying on the principle of acoustic wave superposition [15]. The use of microphone arrays allows to consider the dynamic auditory cues that correspond to the head movements [16]. The use of HRTF datasets allows to consider the individual auditory cues that arises from the interactions of sound waves with the external anatomical shapes of the listeners [14].

Spherical distributions of microphones and HRTF positions are of particular interest because they can sense individual auditory spaces in all directions. Uniform spherical distributions further enable the array signal processing in a transform domain provided by the spherical Fourier transform (SFT) [17]. The orthonormal basis functions involved in the SFT are in turn solutions to the acoustic wave equation at scalable spatial resolutions [18]. The SFT provides a flexible and scalable framework for the processing of spatial sound information. This processing framework is hereafter referred to as transform-domain acoustics.

Conventional transform-domain acoustics methods for binaural systems have hitherto been focused on the reconstruction of acoustic environments by solely considering the recording of spatial sound for its direct binaural synthesis [7–13]. However, little attention has been given to the provision of spatial edition capabilities in between the recording and binaural synthesis stages. Although two studies have addressed separately the variation of sound source distances [19, 20] and the discrimination of sounds along directions [21, 22], to the extent of the authors' knowledge no comprehensive treatment of spatial recording, spatial edition, and binaural synthesis has been formulated so far.

A unified formulation of spatial recording, spatial edition, and binaural synthesis would be useful, for example, to enable the development of the binaural system outlined in Fig. 1. Such binaural system would be capable of recording far sounds in all directions. The recorded far sounds can subsequently be discriminated along directions, and independently brought closer to the listener. The signals obtained during binaural synthesis would correspond to sounds that would be heard as if they were closer to the listener than their original far distances.

To formulated such a unified model, a clear understanding of spatial editing operations (i.e., angle discrimination and distance variation), in both the spatial domain and its transform domain,



$2a$
Microphone array recordings
$\mathcal{P}$
$2b$
HRTF dataset
$\mathcal{H}$
Array signal processing based on transform-domain acoustics
• Combination of array recordings and HRTF datasets
• Inclusion of spatial editing operations.
$\{\mathcal{W}_\ell, r_\ell\}_{\ell=1,\ldots,L}.$
Synthesized binaural signals (to presentation system)
$\mathcal{B}$
$\mathcal{W}_1$
r_1
r_2
$\mathcal{W}_2$
r_L
$\mathcal{W}_L$

Fig. 1 Concept proposal for a binaural system capable of performing spatial edition operations such as the discrimination of sound sources along directions and the manipulation of sound source distances. Angular windows $\mathcal{W}_\ell$ are used to discriminate far sounds along angular regions. Discriminated far sounds can be approached to desired distances r_ℓ.

is crucial for the orderly inclusion of these operations in existing binaural synthesis methods that rely on combining array recordings and HRTF datasets. A body of transform-domain acoustics methods systematically reviewed in [23] sheds lights in this regard to clarify that the spatial operations need to be applied in the following order. First, distance variations are applied independently to the array recordings and HRTF datasets. Second, angular discriminations are applied when combining the distance-edited array recordings and HRTF datasets relying on the principle of acoustic wave superposition.

Bearing the above considerations in mind, the remainder of this paper presents a unified model for binaural systems in spherical geometries to enable the development of the concept outlined in Fig. 1. The model is represented in both the spatial and transform domains. The main difference between these two representations is the method used to apply the angular windows during acoustic wave superposition. It will be shown that the transform-domain method is more exact than the spatial-domain method because additional schemes for numerical integration are required when performing acoustic superposition in the spatial-domain.

2. Formulation overview

Figure 2 shows an overview of the model formulation. The left panel shows the spherical cordinate system used throughout this paper. In the next panels, the spatial sound recordings $\mathcal{P}$ are obtained over a rigid spherical surface composed of recording points $\vec{a}$. The HRTF datasets $\mathcal{H}$ are obtained for a continuous spherical distribution composed of radiating points $\vec{b}$. The binaural signals $\mathcal{B}$ are synthesized in two steps respectively described in the middle and right panels. Operations will involve the spatial and transform domains.

Let $\mathcal{F}$ be a general function in space that is square-integrable on the unit sphere $\mathbb{S}^2$ (i.e., $\mathcal{F}$ can represent the recordings $\mathcal{P}$ or the HRTFs datasets $\mathcal{H}$). The spherical Fourier transform of $\mathcal{F}$ is defined by [17]:

$$\mathcal{F}_{nm}(r) = \mathcal{S}\left\{\mathcal{F}\right\} = \int_{\Omega\in\mathbb{S}^2} \mathcal{F}(r,\Omega)\overline{Y_{nm}(\Omega)}d\Omega, \qquad (1)$$

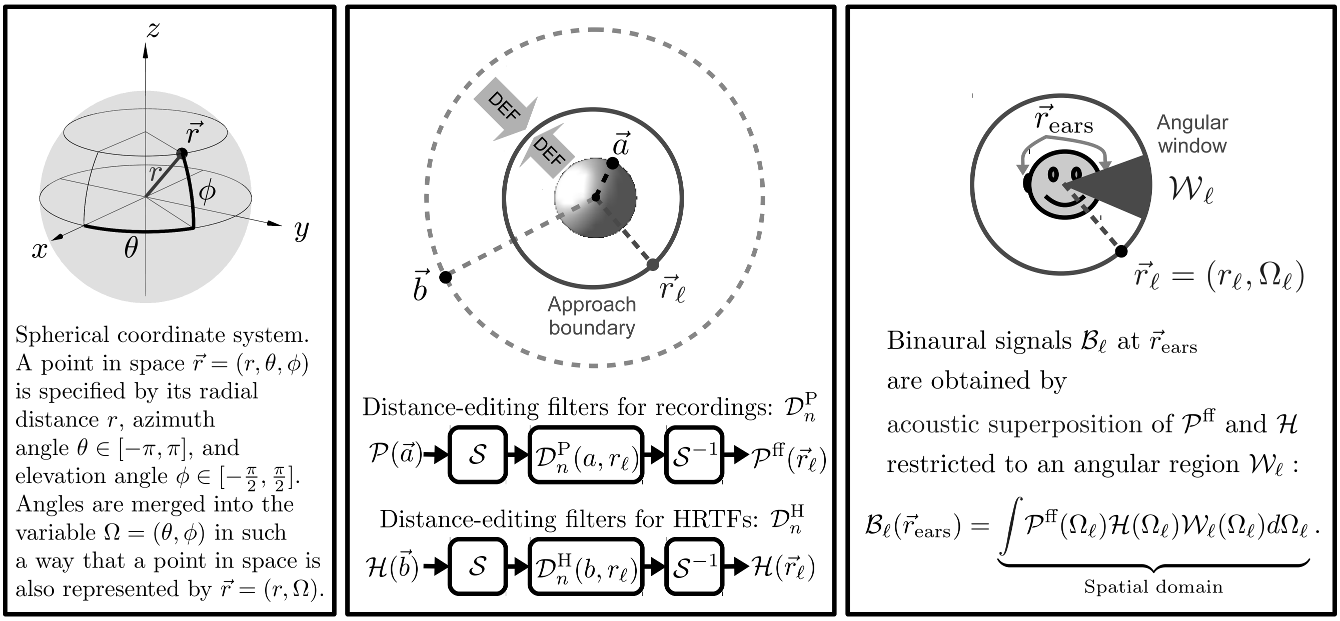
where Y_{nm} are the orthonormal basis functions on the sphere known as the spherical harmonic functions of order n and degree m, and the overbar denotes complex conjugate. The inverse spherical Fourier transform of the expansion coefficients $\mathcal{F}_{nm}$ is defined by [17]:

$$\mathcal{F}(r,\Omega) = \mathcal{S}^{-1}\left\{\mathcal{F}_{nm}\right\} = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} \mathcal{F}_{nm}(r)Y_{nm}(\Omega). \qquad (2)$$

In the middle panel of Fig. 2, distance edition operations are independently applied to the array recordings $\mathcal{P}(\vec{a})$ and HRTF datasets $\mathcal{H}(\vec{b})$ in the transform domain. The distance-editing filters $\mathcal{D}_n^{\mathrm{P}}(a, r_\ell)$ are applied to $\mathcal{P}(\vec{a})$ in the transform domain so as to calculate a free-field distribution of pressure $\mathcal{P}^{\mathrm{ff}}(\vec{r}_\ell)$. These filters are defined as follows [18]:

$$\mathcal{D}_n^{\mathrm{P}}(a, r_\ell) = \frac{-ka^2 h_n'(ka)}{h_n(kr_\ell)}, \qquad (3)$$

where h_n are the spherical Hankel Functions of order n and the symbol $'$ indicates derivative with respect to the argument. The filters in (3) removes the effects of the rigid sphere and further extrapolates the pressure signals from a to r_ℓ. The distance-editing



z
$\vec{r}$
r
ϕ
x
θ
y
Spherical coordinate system. A point in space $\vec{r} = (r, \theta, \phi)$ is specified by its radial distance r, azimuth angle $\theta \in [-\pi, \pi]$, and elevation angle $\phi \in [-\frac{\pi}{2}, \frac{\pi}{2}]$. Angles are merged into the variable $\Omega = (\theta, \phi)$ in such a way that a point in space is also represented by $\vec{r} = (r, \Omega)$.
DEF
DEF
$\vec{a}$
$\vec{b}$
$\vec{r}_\ell$
Approach boundary
Distance-editing filters for recordings: $\mathcal{D}_n^{\mathrm{P}}$
$\mathcal{P}(\vec{a}) \rightarrow \boxed{\mathcal{S}} \rightarrow \boxed{\mathcal{D}_n^{\mathrm{P}}(a, r_\ell)} \rightarrow \boxed{\mathcal{S}^{-1}} \rightarrow \mathcal{P}^{\mathrm{ff}}(\vec{r}_\ell)$
Distance-editing filters for HRTFs: $\mathcal{D}_n^{\mathrm{H}}$
$\mathcal{H}(\vec{b}) \rightarrow \boxed{\mathcal{S}} \rightarrow \boxed{\mathcal{D}_n^{\mathrm{H}}(b, r_\ell)} \rightarrow \boxed{\mathcal{S}^{-1}} \rightarrow \mathcal{H}(\vec{r}_\ell)$
$\vec{r}_{\mathrm{ears}}$
Angular window
$\mathcal{W}_\ell$
$\vec{r}_\ell = (r_\ell, \Omega_\ell)$
Binaural signals $\mathcal{B}_\ell$ at $\vec{r}_{\mathrm{ears}}$ are obtained by acoustic superposition of $\mathcal{P}^{\mathrm{ff}}$ and $\mathcal{H}$ restricted to an angular region $\mathcal{W}_\ell$:
$\mathcal{B}_\ell(\vec{r}_{\mathrm{ears}}) = \underbrace{\int \mathcal{P}^{\mathrm{ff}}(\Omega_\ell)\mathcal{H}(\Omega_\ell)\mathcal{W}_\ell(\Omega_\ell)d\Omega_\ell}_{\text{Spatial domain}} .$

Fig. 2 Formulation overview. DEF stands for distance-editing filter.

filters $\mathcal{D}_n^{\mathrm{H}}(b, r_\ell)$, on the other hand, are applied to $\mathcal{H}(\vec{b})$ in the transform domain so as to calculate the near-distance dataset $\mathcal{H}(\vec{r}_\ell)$. These filters are defined as follows [18]:

$$\mathcal{D}_n^{\mathrm{H}}(b, r_\ell) = \frac{h_n(kr_\ell)}{h_n(kb)}. \tag{4}$$

In the right panel of Fig. 2, angle discrimination operations are applied during binaural synthesis. The distance-edited free-field recordings $\mathcal{P}^{\mathrm{ff}}(\vec{r}_\ell)$ are combined with the distance-edited HRTF datasets $\mathcal{H}(\vec{r}_\ell)$ based on acoustic wave superposition in the spatial domain. This combination yields the binaural signals $\mathcal{B}_\ell$ due to sounds confined to an angular region restricted by the angular window $\mathcal{W}_\ell$ and approached to a distance r_ℓ.

Finally, the binaural signals $\mathcal{B}_\ell$ due to a set of angular windows and approach distances $\{\mathcal{W}_\ell, r_\ell\}_{\ell=1,\ldots,L}$ can be combined into a single pair of binaural signals $\mathcal{B}$ by a simple addition owing again to acoustic superposition.

3. Model for binaural systems

This section presents the model in the spatial and transform domains. The main difference between these two representations is the method used to apply the angular windows during acoustic wave superposition.

The model based on spatial-domain acoustic wave superposition stems from the direct implementation of the formulation described in section 2., which leads to the processing structure in the spatial domain shown in the top panel of Fig. 3. Combining distance-edited array recordings and HRTF datasets based on acoustic superposition in the spatial domain, however, necessarily involves the use of additional spherical sampling schemes $\{\Omega_\ell\}_{\ell=1,\ldots,L}$ for numerical integration on the sphere.

The necessity of additional sampling schemes might be overcome by noting that the free-field distribution of pressure $\mathcal{P}^{\mathrm{ff}}(\vec{r}_\ell)$ and the HRTF dataset $\mathcal{H}(\vec{r}_\ell)$ are already available in the transform domain as $\mathcal{P}_{nm}^{\mathrm{ff}}(r_\ell)$ and $\mathcal{H}_{n'm'}(r_\ell)$, respectively. This can be observed in the top panel of Fig. 3. It would be useful, then, if a similar procedure for performing the superposition was available for direct calculation in the transform domain.

A transform-domain procedure for acoustic wave superposition can indeed be obtained by expressing each factor in the integral in terms of their Fourier coefficients using (2). It can be shown that the integral used in the principle of acoustic wave superposition can equivalently be calculated in the transform domain as follows:

$$\begin{aligned}\mathcal{B}_\ell(\vec{r}_{\mathrm{ears}}) &= \int \mathcal{P}^{\mathrm{ff}}(\Omega_\ell)\mathcal{H}(\Omega_\ell)\mathcal{W}(\Omega_\ell)d\Omega_\ell \\ &= \sum_{nm}^{\infty}\sum_{n'm'}^{\infty}\sum_{n''m''}^{\infty} \mathcal{P}_{nm}^{\mathrm{ff}}\mathcal{H}_{n'm'}\mathcal{W}_{\ell,n''m''}\gamma_{nn'n''}^{mm'm''},\end{aligned} \tag{5}$$

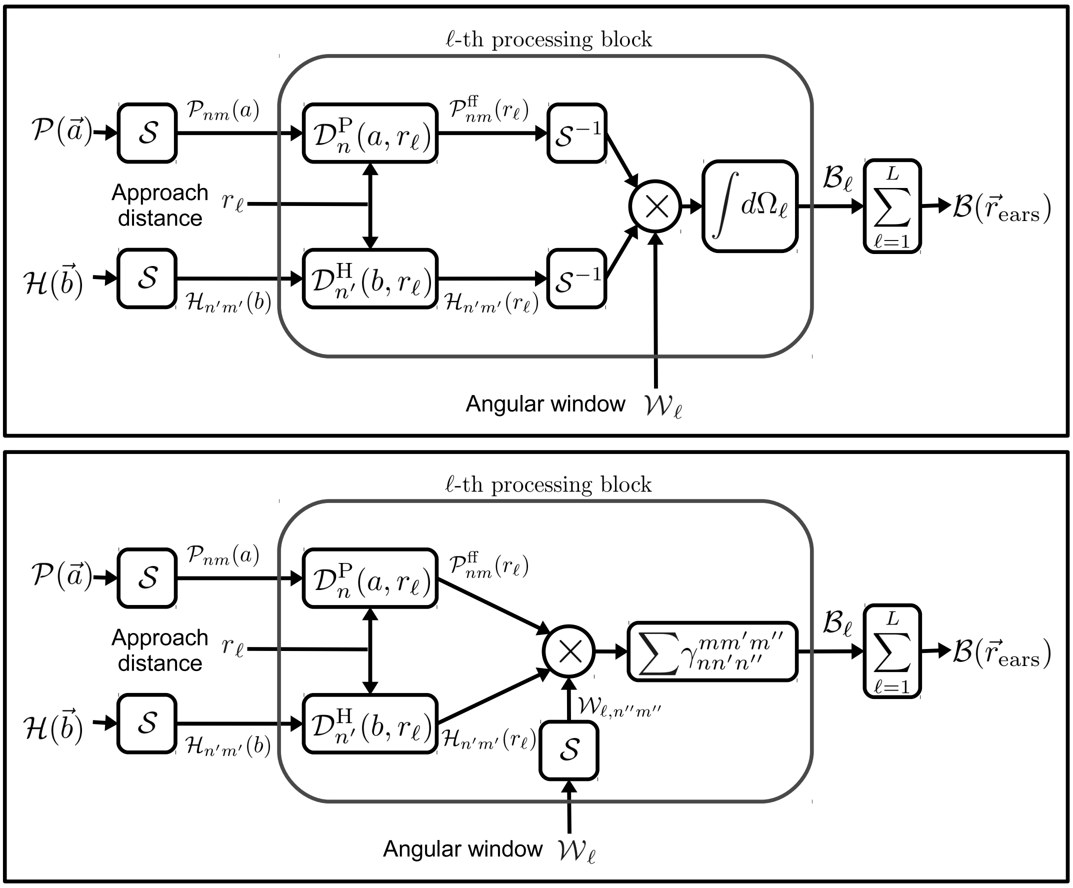
where

$$\gamma_{nn'n''}^{mm'm''} = \int Y_n^m(\Omega)Y_{n'}^{m'}(\Omega)Y_{n''}^{m''}(\Omega)d\Omega. \tag{6}$$

Because analytical expressions exist for the calculation of (6) (see e.g. [24]), no additional sampling schemes are required to calculate the sum in (5). The direct implementation of the model based on the sum in (5) results in the spatial processing structure shown in the bottom panel of Fig.3

4. Examples of binaural synthesis

Binaural signals synthesized with the proposed transform-domain model are exemplified in Fig. 4. Synthesis was performed by assuming a rigid spherical array composed of 252 microphones and a HRTF dataset for 252 positions. The positioning of both



ℓ-th processing block
$\mathcal{P}(\vec{a})$
$\mathcal{S}$
$\mathcal{P}_{nm}(a)$
$\mathcal{D}_n^{\mathrm{P}}(a, r_\ell)$
$\mathcal{P}_{nm}^{\mathrm{ff}}(r_\ell)$
$\mathcal{S}^{-1}$
Approach distance
r_ℓ
$\int d\Omega_\ell$
$\mathcal{B}_\ell$
$\sum_{\ell=1}^{L}$
$\mathcal{B}(\vec{r}_{\mathrm{ears}})$
$\mathcal{H}(\vec{b})$
$\mathcal{H}_{n'm'}(b)$
$\mathcal{D}_{n'}^{\mathrm{H}}(b, r_\ell)$
$\mathcal{H}_{n'm'}(r_\ell)$
Angular window $\mathcal{W}_\ell$
$\sum \gamma_{nn'n''}^{mm'm''}$
$\mathcal{W}_{\ell,n''m''}$

Fig. 3 Models with acoustic wave superposition in the spatial (top) and transform (bottom) domains. In continuous geometries, the models are equivalent. In discrete geometries, however, spatial-domain superposition requires to decide a sampling scheme $\{\Omega_\ell\}_{\ell=1,\ldots,L}$ for numerical integration on the sphere. Errors due to numerical integration are overcome in transform-domain superposition because there exist exact formulas to calculate the coefficients $\gamma_{nn'n''}^{mm'm''}$.

microphones and HRTFs was decided according to spherical grids that are based on the geometry of the icosahedron.

The top-right panel in Fig. 4 shows the binaural signals synthesized with two angular regions discriminated by using a spherical cap window (spherical version of a box window). The approach distances were also different along each angular window. It can be observed that angular regions can be discriminated and the distance patterns in the binaural signals are also properly synthesized. However, some distortions are also observed in these patterns.

The bottom-right panel of Fig. 4 shows the binaural signals synthesized by using a spherical Gaussian window. Synthesis results were smother than the case of a box window. However, in both cases, acceptable synthesis was only possible up to a maximum frequency determined by the number of microphones in the array.

5. Conclusion

A general binaural synthesis model for far and near distances has been introduced. The model can combine and edit distinct types of spatial sound information on the sphere, such as microphone array recordings, acoustic transfer functions, and angular windows. The model constitutes a theoretical basis for the future development of binaural systems capable of recording far sounds in all directions, discriminating sounds in distinct directions, and bringing the selected sounds closer to the listener.

6. Acknowledgments

A part of this study was supported by the JSPS Grant-in-Aid for Scientific Research no. 24240016 and no. 16H01736, and by the A3 Foresight Program for ”Ultra-realistic acoustic interactive communication on next-generation Internet”.

References

[1] Y. Suzuki, T. Okamoto, J. Trevino, Z.-L. Cui, Y. Iwaya, S. Sakamoto, and M. Otani, “3d spatial sound systems compatible with human’s active listening to realize rich high-level *kansei* information,” *Interdiscipl. Inform. Sci.*, vol. 18, no. 2, pp. 71–82, 2012.

[2] A. Steed and R. Schroeder, “Collaboration in immersive and non-immersive virtual environments,” in *Immersed in Media: Telepresence Theory, Measurement & Technology*, Matthew Lombard, Frank Biocca, Jonathan Freeman, Wijnand IJsselsteijn, and J. Rachel Schaevitz, Eds., pp. 263–282. Springer International Publishing, Cham, 2015.

[3] H. Moller, “Fundamentals of binaural technology,” *Appl. Acoust.*,

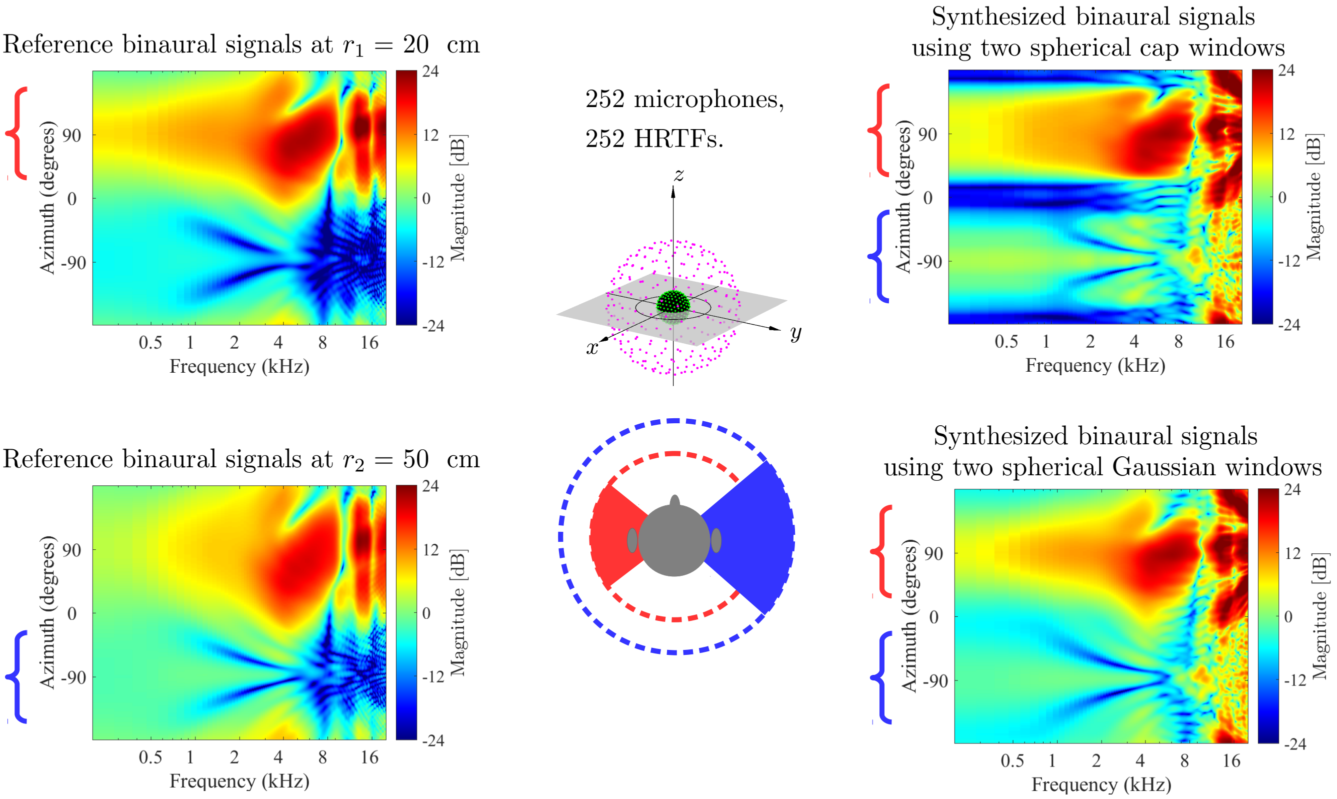


Fig. 4 Examples of binaural synthesis on the horizontal plane using a spherical cap window (top right) and a Gaussian window (bottom right) for the discrimination of sounds along directions.

vol. 36, no. 3–4, pp. 171–218, 1992.

[4] M. Morimoto and Y. Ando, "On the simulation of sound localization," *J. Acoust. Soc. Jpn. (E)*, vol. 1, no. 3, pp. 167–174, 1980.

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

[6] 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.

[7] 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.

[8] R. Duraiswami, D. N. Zotkin, Z. Li, E. Grassi, N. A. Gumerov, and Larry 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.

[9] 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, Sept. 2014.

[10] 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.

[11] 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, Aug. 2016.

[12] C. D. Salvador, S. Sakamoto, J. Trevino, and Y. Suzuki, "Spatial accuracy of binaural synthesis from rigid spherical microphone array recordings," *Acoust. Sci. Technol.*, vol. 38, no. 1, pp. 23–30, 2017.

[13] C. D. Salvador, S. Sakamoto, J. Trevino, and Y. Suzuki, "Design theory for binaural synthesis: combining microphone array recordings and HRTF datasets," *Acoust. Sci. Technol.*, vol. 38, no. 2, 2017, in print.

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

[15] 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.*, vol. 86, no. 6, pp. 2433–2438, 1989.

[16] 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.

[17] J. R. Driscoll and D. M. Heally, "Computing Fourier transforms and convolutions on the 2-sphere," *Adv. Appl. Math.*, vol. 15, pp. 202–250, 1994.

[18] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography*, Academic Press, London, UK, 1999.

[19] C. D. Salvador, S. Sakamoto, J. Trevino, and Y. Suzuki, "Editing distance information in compact microphone array recordings for its binaural rendering," *IEICE technical report*, vol. 114, no. 3, pp. 13–18, 2014.

[20] C. D. Salvador, S. Sakamoto, J. Trevino, and Y. Suzuki, "Embedding distance information in binaural renderings of far field recordings," in *Proceedings of the EAA Joint Symposium on Auralization and Ambisonics*, Berlin, Germany, Apr. 2014, pp. 133–139.

[21] N. R. Shabtai and B. Rafaely, "Generalized spherical array beamforming for binaural speech reproduction," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 1, pp. 238–247, Jan. 2014.

[22] N. R. Shabtai, "Optimization of the directivity in binaural sound reproduction beamforming," *J. Acoust. Soc. Am.*, vol. 138, no. 5, pp. 3118–3128, 2015.

[23] C. D. Salvador, *Binaural Synthesis Based on Spherical Acoustics*, Doctoral dissertation, Tohoku University, Sendai, Sept. 2016.

[24] B. C. Carlson and G. S. Rushbrooke, "On the expansion of a Coulomb potential in spherical harmonics," *Math. Proc. Cambridge Philos. Soc.*, vol. 46, no. 4, pp. 626–633, 1950.