Acoust. Sci. & Tech. **38**, 1 (2017)

TECHNICAL REPORT

Spatial accuracy of binaural synthesis from rigid spherical microphone array recordings

César D. Salvador*, Shuichi Sakamoto†, Jorge Treviño‡ and Yôiti Suzuki§

Graduate School of Information Sciences and Research Institute of Electrical Communication, Tohoku University,
2–1–1 Katahira, Aoba-ku, Sendai, Miyagi, 980–8577 Japan

(*Received 15 April 2016, Accepted for publication 25 July 2016*)

Abstract: Binaural systems are a promising class of three-dimensional (3D) auditory displays for high-definition personal 3D audio devices. They properly synthesize the sound pressure signals at the ears of a listener, namely binaural signals, by means of the head-related transfer functions (HRTFs). Rigid spherical microphone arrays (RSMAs) are widely used to capture sound pressure fields for binaural presentation to multiple listeners. However, the spatial resolution needed in the RSMAs to allow for accurate binaural reproduction has not been studied in detail. The aim of this paper is to objectively address this question. We evaluated the spatial accuracy in binaural signals synthesized from the recordings of RSMAs with different number of microphones using the model of a human head. We find that the synthesis of spectral cues is accurate up to a maximum frequency determined by the number of microphones. Nevertheless, we also identify a limit beyond which adding more microphones does not improve overall accuracy. Said limit is higher for the interaural spectral cues than for the monaural ones.

Keywords: 3D auditory displays, Binaural technology, Head-related transfer functions, Spherical microphone array, Spherical acoustics

PACS number: 43.60.Fg, 43.60.Uv, 43.66.Pn [doi:10.1250/ast.38.23]

1. INTRODUCTION

Three-dimensional (3D) auditory displays are a key component to realize future affective multimodal communications because the spatial features of sound can enhance the semantic and emotional cues perceived by the listeners [1]. The aim of 3D auditory displays is to provide the listeners with the sense of being present in a virtually rendered or reproduced acoustic environment. The appropriate synthesis of the spatial features of sound is hence important to this end.

Binaural systems [2–4] are a promising class of 3D auditory displays, specially for personal high-definition audio devices, because of their small size, simple signal processing, and good performance. Binaural systems aim to synthesize the sound pressure signals that would be generated at the ears of each listener, namely *binaural signals*, as if the listeners were present in a specific acoustic environment.

*e-mail: salvador@ais.riec.tohoku.ac.jp
†e-mail: saka@ais.riec.tohoku.ac.jp
‡e-mail: jorge@ais.riec.tohoku.ac.jp
§e-mail: yoh@riec.tohoku.ac.jp

Synthesizing the binaural signals for a position in a specific acoustic environment would require to obtain the transfer functions of the sound propagation paths from that position to the ears of the listeners. The binaural signals can subsequently be generated by multiplying such acoustic transfer functions with a sound source signal recorded in free-field conditions. The acoustic transfer functions can be considered to be composed of a representation of the environment (e.g. room acoustics) and a representation of the coloration caused by the interactions of incident sound with the listener's pinna, head and body [5]. The latter component is called the head-related transfer function (HRTF) and is a function of not only frequency but also of the sound source position relative to the center of the listener's head. Moreover, HRTFs are highly individual because they are dependent on each listeners' external anatomical shapes of and around the ears. Individual HRTFs, therefore, comprehensively involve perceptual sound localization cues when listening is done in free field conditions.

Binaural systems require sets of individual HRTFs characterized for densely distributed sound sources to enable the auditory localization at any specified position.

Such HRTF datasets are typically obtained for a spherical array of sound sources at a single distance from the listener's head [6]. Because HRTFs hardly depend on distance when sources are beyond 1 to 1.5 m [7–9], the radius of the array is typically taken within or beyond this interval.

The recording of spatial sound with almost uniform radial symmetry, on the other hand, is possible with a microphone array mounted on the surface of a rigid sphere. This kind of array is called a rigid spherical microphone array (RSMA). To enable binaural reproduction, the RSMA recordings need to be adapted for its rendering with HRTF datasets.

During the last decade, there has been an increasing interest on methods for combining spatial information contained in HRTF datasets with recordings made with RSMAs [10–20]. In particular, the use of representations of HRTF datasets and RSMA recordings, in terms of solutions to the acoustic wave equation at different spatial resolutions (orders), enjoy popularity because they enable scalable encoding and simplify multichannel processing [10,11,14–16,18,19].

HRTF datasets can be obtained for high-resolution source distributions using numerical methods [21]. Recently, a perceptual study [22] has reported that low-order HRTF representations might be sufficient to approximate an individual space, whereas an objective study [23] has identified HRTF features that would require high-order representations. Because the required resolution for characterizing an individual space is still an open question, the performance of binaural systems should be evaluated by considering HRTF datasets with the higher resolution that can be achieved.

Existing RSMA recordings, on the other hand, typically contain low-order information only, since high-resolution RSMAs are still hard to construct using actual technology. This has motivated a recent study [16] on adapting the resolutions of HRTF datasets to RSMA recordings by spatially downsampling the HRTF datasets. Nevertheless, it has also recently been reported in [15] that high-order information is important to synthesize more directionally sharpened and more externalized (outside-head) sounds. Regarding the future of recording technology, RSMAs of hundreds of microphones are not unrealistic. For instance, it has been reported in [12,17,20] a setup composed of 252 microphones distributed according to an icosahedral symmetry. Plans to construct higher resolution arrays in the near future also exist.

Bearing these considerations in mind, the present study seek to identify the number of microphones that are necessary to synthesize the binaural signals with a specified spatial accuracy. We therefore present an extensive numerical evaluation using RSMA recordings and HRTF datasets of different resolutions, up to the amount required to cover all typical audible frequencies objectively. To focus the analysis on the number of microphones, spatial encodings of RSMA recordings are calculated only. In connection with [16], this is equivalent to a spatial resampling of the RSMA recordings so as to match the resolution of the HRTF datasets. Furthermore, to cope with low-frequency high amplifications typically observed when assuming HRTF datasets characterized for plane-wave sources, we consider the case of point sources, which is more consistent with existing HRTF datasets. An overview of the binaural system under consideration is shown in Fig. 1.

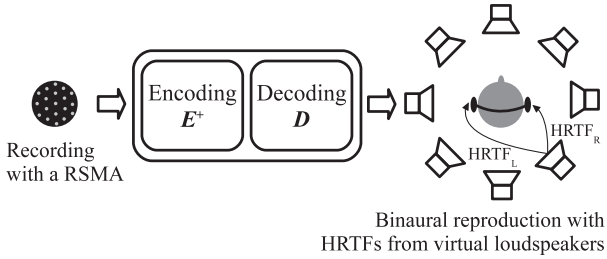


Fig. 1 Overview of binaural systems composed of RSMA recording, spatial encoding and decoding, and binaural reproduction.

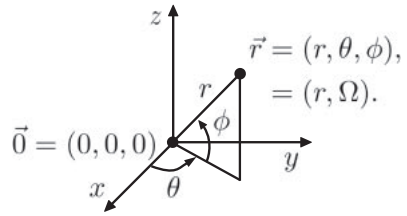


Fig. 2 Spherical coordinate system. The origin $\vec{0}$ coincides with the center of the array and the center of the listener's head.

The reminder of this paper is structured as follows. Section 2 presents a discrete spatial model of the binaural system under consideration. Section 3 presents the conditions and results of numerical experiments. Concluding remarks are stated in Sect. 4.

2. BINAURAL SYNTHESIS

In the spherical coordinate system shown in Fig. 2, 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 can be merged into the variable $\Omega = (\theta, \phi)$ in such a way that a point in space is also represented by $\vec{r} = (r, \Omega)$.

In the frequency domain, we define an input vector

$$\boldsymbol{p} = \begin{bmatrix} p_1 & p_2 & \cdots & p_Q \end{bmatrix}^\top. \tag{1}$$

The symbol $^\top$ indicates transpose. The entries of $\boldsymbol{p}$ are pressure signals p_q recorded at Q microphone positions $\vec{r}_q^{\,\mathrm{m}} = (r_\mathrm{m}, \Omega_q^\mathrm{m})$ over a rigid baffle of radius r_m, where superscript or subscript 'm' refers to the microphones, and $q = 1, 2, \ldots, Q$.

We also define a user parameter matrix

$$\boldsymbol{h} = \begin{bmatrix} h_1^\mathrm{left} & h_2^\mathrm{left} & \cdots & h_L^\mathrm{left} \\ h_1^\mathrm{right} & h_2^\mathrm{right} & \cdots & h_L^\mathrm{right} \end{bmatrix}^\top \tag{2}$$

It contains the free-field HRTFs for the left and right ears, respectively denoted by h_ℓ^left and h_ℓ^right. They are characterized for L positions $\vec{r}_\ell^{\,\mathrm{v}} = (r_\mathrm{v}, \Omega_\ell^\mathrm{v})$ on a radius r_v, where $\ell = 1, 2, \ldots, L$. These positions are called the *virtual loudspeaker* positions, and superscript or subscript 'v' is used for referring to them.

The synthesized *binaural signals* for the left and right ears are organized in the pair

$$\hat{\boldsymbol{b}} = \begin{bmatrix} \hat{b}^\mathrm{left} & \hat{b}^\mathrm{right} \end{bmatrix}^\top. \tag{3}$$

Binaural synthesis can be summarized as the following linear combination of $\boldsymbol{p}$ and $\boldsymbol{h}$:

$$\hat{\boldsymbol{b}} = \boldsymbol{h}^\top \boldsymbol{A} \boldsymbol{p}. \tag{4}$$

Here, the combination matrix of size $L \times Q$ is calculated according to the following expression:

$$\boldsymbol{A} = \boldsymbol{D}\boldsymbol{E}^+, \tag{5}$$

where $\boldsymbol{E}^+$ and $\boldsymbol{D}$ denote the encoding and decoding matrices, respectively. In general terms, the matrix $\boldsymbol{A}$ can also be interpreted as a spatial resampling from the resolution of the RSMA to the resolution of the HRTF dataset.

The matrix $\boldsymbol{E}^+$ calculates a representation of $\boldsymbol{p}$ in terms of harmonic solutions to the acoustic wave equation up to order N. It further compensates for the presence of the spherical baffle, and extrapolates the resulting free-field representation from r_m to r_v. The matrix $\boldsymbol{E}^+$ is obtained by calculating the pseudo-inverse of a matrix $\boldsymbol{E}$ by using Tikhonov regularization [24]. The entries of $\boldsymbol{E}$ represent acoustic transfer functions from an arbitrary position at a distance r_v to each microphone position $\boldsymbol{r}_q^\mathrm{m}$. The size of $\boldsymbol{E}$ is $Q \times (N+1)^2$ and its entries, for the assumption of virtual loudspeakers radiating spherical waves, are

$$e_{q,n^2+n+m+1} = \frac{-h_n(kr_\mathrm{v}) Y_n^m(\Omega_q^\mathrm{m})}{kr_\mathrm{m}^2 h'_n(kr_\mathrm{m})}. \tag{6}$$

Here, h_n denotes the spherical Hankel function of the second kind and order n, while Y_n^m denote the complex spherical harmonic functions of order n and degree m, where $n = 0, 1, \ldots, N$, and $m = -n, -n+1, \ldots, n$. The functions h_n and Y_n^m are defined in [25], respectively as the radial and angular portions of the solutions to the acoustic wave equation. They are also functions of the wave number $k = \frac{2\pi f}{c}$, where f denotes frequency and c denotes the speed of sound in air. The symbol $'$ denotes the derivative of a function with respect to its argument. The benefit of performing regularization by including the radial portion is the attenuation of high-order components at low frequencies.

The matrix $\boldsymbol{D}$ takes the encodings $\boldsymbol{E}^+\boldsymbol{p}$ at a radius r_v and decodes them for each virtual loudspeaker directions Ω_ℓ^v. The size of $\boldsymbol{D}$ is $L \times (N+1)^2$ and its entries are

$$d_{\ell,n^2+n+m+1} = \frac{\exp(jkr_\mathrm{v})}{r_\mathrm{v}} Y_n^m(\Omega_\ell^\mathrm{v}). \tag{7}$$

The fraction represents a free-field transfer function from r_v to the head center. Its purpose is to set the head center as the observation point, in consistency with the definition of free-field HRTFs.

3. SPATIAL ACCURACY

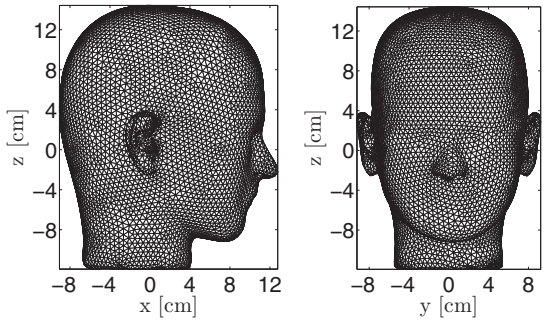
We evaluated the effect of using different numbers of microphones (Q) and virtual loudspeakers (L) on the synthesis accuracy. To emphasize the preservation of spectral information used in human auditory localization, we gave special attention to the synthesis of monaural and interaural spectral cues.

Only one example virtual loudspeaker radius $r_\mathrm{v} = 1.5$ m was used. Most of the available HRTF datasets are measured at this typical radius, beyond which the HRTFs hardly depend on distance [9]. The sound source to be recorded was also assumed placed at a 1.5 m distance. An exhaustive evaluation at different distances close to the head, while important, is outside the intended scope of this paper. We also leave as future work the evaluations using a signal model with noise.

3.1. Conditions of the Evaluation

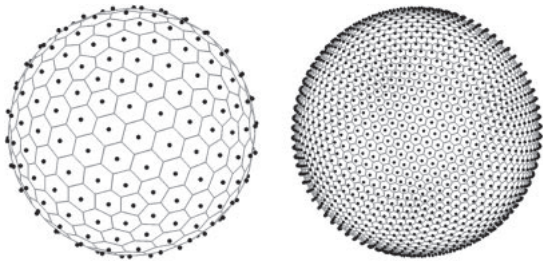
By $\hat{B} = \{\hat{B}_\mathrm{left}(\Omega_i, f_j), \hat{B}_\mathrm{right}(\Omega_i, f_j)\}$, we denote a set of *binaural transfer functions*, where $i = 1, 2, \ldots, I$, and $j = 1, 2, \ldots, J$. The set $\hat{B}$ was calculated using (4) for the particular case of a number I of point sources placed at a 1.5 m distance, in the directions $\Omega_i = (\theta_i, \phi_i)$. The point sources were radiating non-simultaneously. Each point source was radiating a single sinusoidal signal at a time, and this case was repeated for a number J of single frequencies f_j in the full audible range. Similarly, by $H = \{H_\mathrm{left}(\Omega_i, f_j), H_\mathrm{right}(\Omega_i, f_j)\}$, we denote the reference HRTF datasets (target) calculated using the boundary element method (BEM) [21] for the head model described in Fig. 3. The mesh grid of this head model consists of 14,096 points with an average cell length of 5.1 mm, which limited our evaluations to an average frequency of 16.6 kHz. Evaluations were based on comparisons of $\hat{B}$ and H.

To calculate $\hat{B}$, microphone signals for the non-simultaneous point sources (input) were first calculated



(a) Lateral view. (b) Front view.

Fig. 3 Head model used for numerical experiments.



(a) 252 points. (b) 1962 points.

Fig. 4 Examples of icosahedral grids used for the arrangement of microphones and virtual loudspeakers.

with the algorithm in [26], assuming a rigid sphere of radius $r_\mathrm{m} = 8.5\,\mathrm{cm}$. Then, HRTF datasets for a virtual loudspeaker array of radius $r_\mathrm{v} = 1.5\,\mathrm{m}$ (user parameter) were calculated using BEM [21] and the head model shown in Fig. 3. The positions of microphones and virtual loudspeakers were decided using spherical grids constructed by subdividing the edges of the icosahedron into equal segments. Examples of such grids are shown in Fig. 4, where dots indicate the positions and lines enclose their Voronoi cells. Although not explicitly mentioned in (6) and (7), encoding and decoding fundamentally lie on numerical integrations on the sphere. The required quadrature weights were thus determined to be proportional to the cell areas.

To select the maximum order N_j required to approximate a frequency f_j, we used the bound proposed in [27]. Maximum orders N_j are determined by setting a constant truncation error within a region of interest enclosed by a given radius. In our simulations, we set a truncation error equal to 10^{-5} within the radius $r_\mathrm{m} = 8.5\,\mathrm{cm}$ of the microphone array. Source distance information is also considered by this rule; we set this variable equal to 1.5 m. Under these conditions, approximations up to an average limit frequency of 16.6 kHz would require an order $N = 43$ and, hence, at least $Q = (43+1)^2 = 1{,}936$ microphones. Conversely, a given number of microphones would limit the directional resolution up to a maximum frequency.

The regularization parameter required to calculate $\boldsymbol{E}^+$ was empirically chosen so as to obtain a good compromise between the error and the energy of the source. It was therefore set equal to $\|\boldsymbol{E}\| \times 10^{-7}$, with the norm of $\boldsymbol{E}$ equal to its largest singular value.

3.2. Objective Measures of Accuracy

Sets $\hat{B}$ and H were compared giving special attention to the spectral monaural and interaural localization cues. The monaural local error in decibels is defined by [28]:

$$E_\mathrm{M}(\Omega_i, f_j) = 20\log_{10}\left|\frac{\hat{B}_\mathrm{left}(\Omega_i, f_j)}{H_\mathrm{left}(\Omega_i, f_j)}\right|. \tag{8}$$

To highlight the capability of synthesizing the main peaks and notches of the HRTFs, which provide important features for auditory localization, the overall gain mismatch was further removed from (8) by subtracting its overall mean value $\bar{E}_\mathrm{M}$.

Because interaural information is important in sound localization, as opposed to monoaural phase [29], interaural measures of accuracy were also considered.

The reference interaural HRTFs are defined as [5]

$$H_\mathrm{interaural}(\Omega_i, f_j) = \frac{H_\mathrm{left}(\Omega_i, f_j)}{H_\mathrm{right}(\Omega_i, f_j)}, \tag{9}$$

and the synthesized interaural transfer functions as

$$\hat{B}_\mathrm{interaural}(\Omega_i, f_j) = \frac{\hat{B}_\mathrm{left}(\Omega_i, f_j)}{\hat{B}_\mathrm{right}(\Omega_i, f_j)}. \tag{10}$$

The interaural level differences (ILDs) corresponding to the reference and synthesized transfer functions are defined by the magnitude in decibels of (9) and (10), respectively. We calculated the ILD local error correspondingly in decibels by

$$E_\mathrm{ILD}(\Omega_i, f_j) = 20\log_{10}\left|\frac{\hat{B}_\mathrm{interaural}(\Omega_i, f_j)}{H_\mathrm{interaural}(\Omega_i, f_j)}\right|. \tag{11}$$

On the other hand, the interaural phase differences (IPDs) associated with the reference and synthesized transfer functions correspond to the phase in radians of (9) and (10), respectively. In particular, phase information can be displayed by means of its group delay to highlight spectral information related to peaks and notches with a better resolution [30]. We calculated the interaural group delay (IGD) local error correspondingly in seconds according to

$$E_\mathrm{IGD}(\Omega_i, f_j) = \frac{\Delta\arg\left(\dfrac{\hat{B}_\mathrm{interaural}(\Omega_i, f_j)}{H_\mathrm{interaural}(\Omega_i, f_j)}\right)}{2\pi\Delta f_j}, \tag{12}$$

where arg denotes the unwrapped phase and Δ is the finite difference operator along the discrete variable f_j.

We examined the overall accuracy of our method based on the root mean square (RMS) values of the errors defined

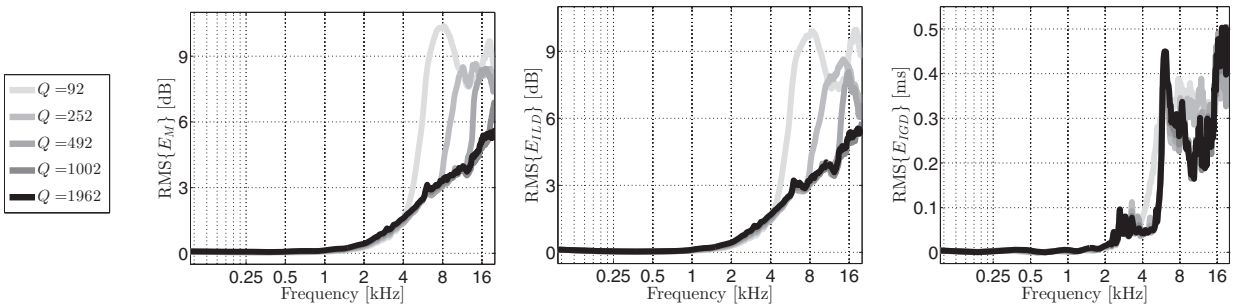


Fig. 5 Overall accuracy for sources on the sphere calculated using (13), for synthesis with $L = 1{,}962$ virtual loudspeakers and different numbers of microphones (Q).

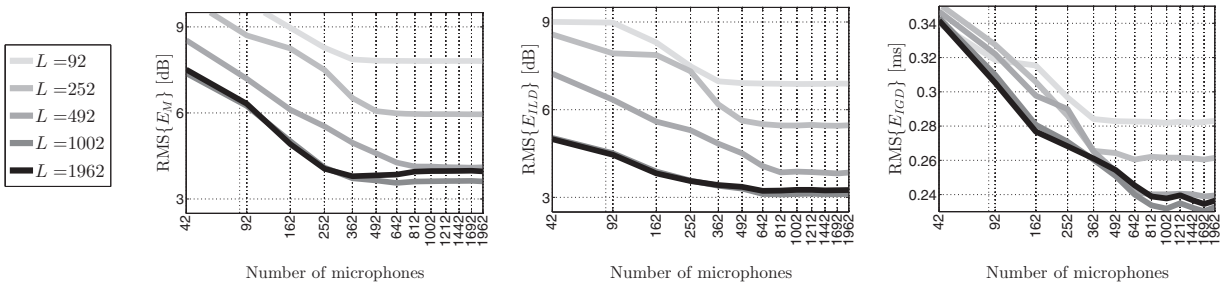


Fig. 6 Overall accuracy for sources on the sphere calculated using (14) up to 16.6 kHz, for synthesis with different numbers of virtual loudspeakers (L).

in (8), (11) and (12). Norms of accuracy equivalent to the RMS value have been evaluated through listening tests in [28], where the suitability of these norms for predicting audible differences between measured and synthesized HRTFs was verified. We calculated the RMS value along directions based on the following expression:

$$\underset{\Omega_i}{\mathrm{RMS}}\{E\}(f_j) = \left(\sum_{i=1}^{I} E(\Omega_i, f_j)^2 w_i\right)^{\frac{1}{2}}. \qquad (13)$$

Here, E can be one of the errors E_M in (8), E_ILD in (11), or E_IGD in (12), and w_i are normalized quadrature weights (area of Voronoi cells) for numerical integration over all of the sound source directions on a sphere with a radius of 1.5 m. Normalization is done in such a way that $\sum_i w_i = 1$, where $w_i > 0$. Similarly, we extended the calculation of the RMS value to cover all directions and frequencies according to:

$$\underset{\Omega_i, f_j}{\mathrm{RMS}}\{E\} = \left(\frac{1}{J}\sum_{i,j=1}^{I,J} E(\Omega_i, f_j)^2 w_i\right)^{\frac{1}{2}}. \qquad (14)$$

3.3. Synthesis on the Sphere

We considered $I = 5{,}762$ sound sources almost uniformly distributed on a sphere with a radius of 1.5 m and frequency bins in the full audible range for a sampling frequency of 48 kHz. The results based on (13) and (14) are displayed in Figs. 5 and 6, respectively.

In Fig. 5, it can be observed that when a number of virtual loudspeakers sufficient to cover the audible frequency range was used, excellent monaural and interaural overall accuracies were obtained up to 2 kHz, even with a limited number of microphones. Over 2 kHz, accuracies of monaural levels and interaural level differences (left and middle panels) gradually decreased with increasing frequency and a decreasing number of microphones. Nevertheless, increasing the number of microphones beyond 1,002 did not yield a significant improvement in these overall accuracies. Regarding the interaural group delay (right panel), good performance was also obtained at low frequencies, precisely where these cues are known to be important.

In Fig. 6, it can be observed that both monaural and interaural accuracies were also degraded when the number of virtual loudspeakers decreased below 1,002. However, increasing this number over 1,002 did not improve the overall accuracies. On the other hand, overall accuracies were significantly improved by increasing the number of microphones up to a certain limit, after which adding more microphones did not lead to a decrease in the RMS errors. For a number of virtual loudspeakers greater than 1,002, the left panel shows that increasing the number of microphones beyond 362 did not improve the overall accuracy of the monaural cues. For the same condition, the middle panel shows that the overall accuracy in ILD did

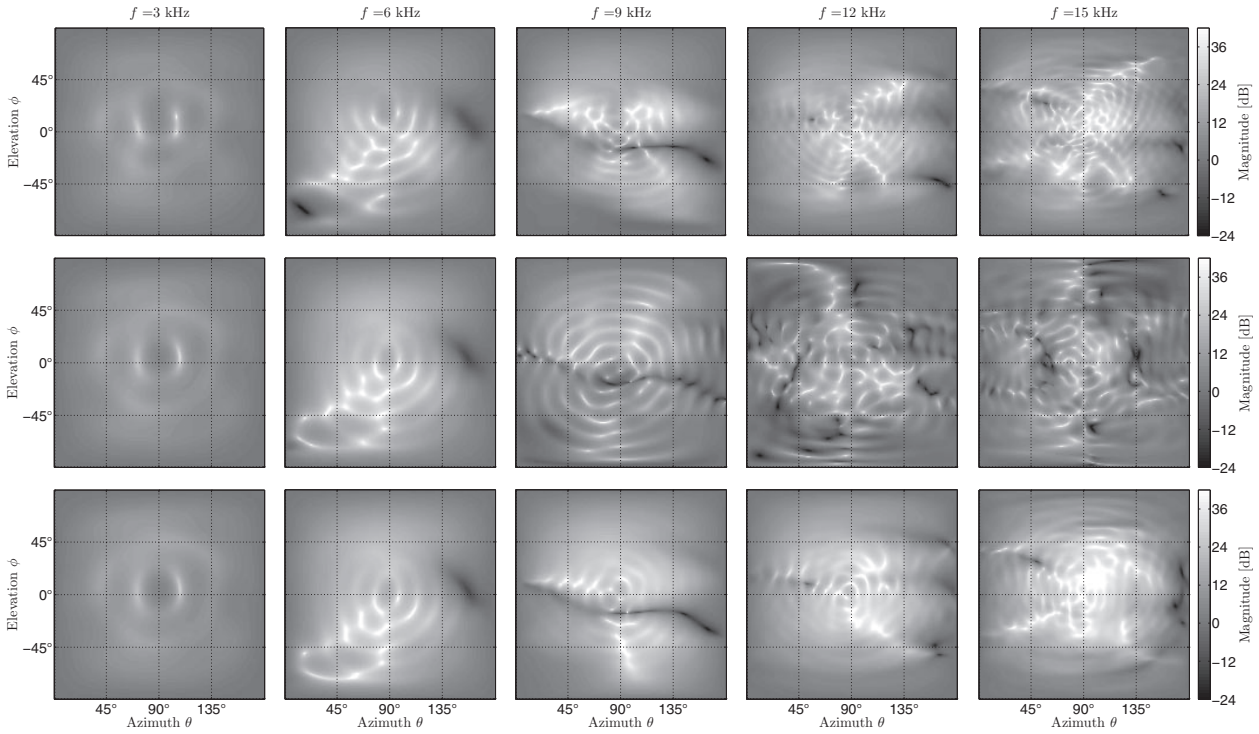


Fig. 7 Interaural level differences (ILDs) on the sphere. Top row shows the reference ILDs. Middle row shows the ILDs synthesized using $Q = 252$ microphones and $L = 1{,}962$ virtual loudspeakers. Bottom row shows the ILDs synthesized using $Q = 1{,}962$ microphones and $L = 1{,}962$ virtual loudspeakers.

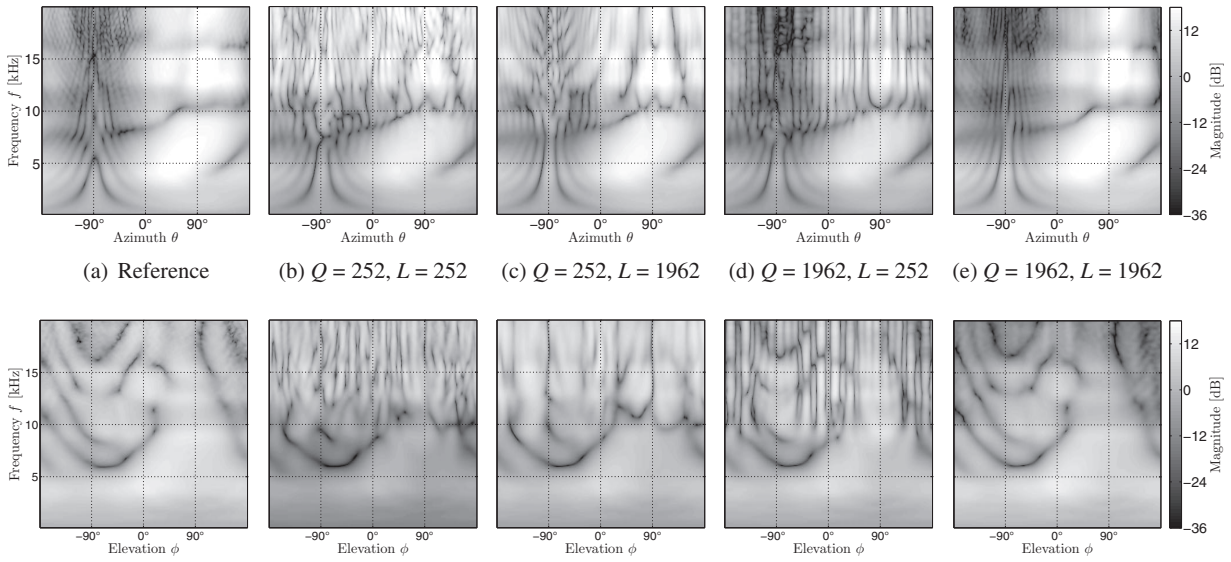
not benefit from additional microphones beyond 642. Furthermore, for more than 1,002 virtual loudspeakers, the right panel shows that increasing the number of microphones beyond around 1,002 did not improve the overall IGD accuracy. The limits were different depending on the type of spectral cue under consideration.

The overall accuracies obtained for 5,762 sources can be used to predict, to some extent, the local accuracy when synthesis is performed for denser distributions of sources on the sphere. To exemplify this, the spectral cues for dense datasets and some single frequencies have been synthesized. A qualitative and visual comparison between the reference and synthesized datasets is briefly described below.

Figure 7 shows some examples of reference (top row) and synthesized ILDs, for 360×180 sound sources equiangularly distributed on the sphere. Given the symmetry of the head model, only the left hemisphere is shown. The middle row shows the synthesized ILDs when the array of 252 microphones available in our institute was assumed [12,17,20]. Good accuracies were obtained for the cases of 3 and 6 kHz, as expected from the curve for $Q = 252$ in the center panel of Fig. 5, where RMS values below 3 dB are observed up to around 8 kHz. The bottom row shows the synthesized ILDs when an array of 1,962 microphones was assumed. In theory, this provides a directional resolution sufficiently high to cover frequencies up to the limit of 16.6 kHz imposed by the head model. In this case, good performance was obtained at 3 and 6 kHz. The performance at 9 kHz was also acceptable. However, local distortions were observed in the ILDs for the cases of 12 and 15 kHz, specially in the lateral region around azimuth $\theta = 90^\circ$ and elevation $\phi = 0^\circ$.

3.4. Synthesis on the Horizontal and Median Planes

Figure 8 shows the reference and synthesized HRTFs for the left ear and $I = 360$ sound sources equiangularly distributed on the horizontal plane (top row) and the median plane (bottom row). A visual comparison with the reference HRTFs in the first column shows that transfer functions in the second column, synthesized with $Q = 252$ microphones and $L = 252$ virtual loudspeakers, show fair agreement up to around 10 kHz. However, a detailed inspection of the horizontal plane shows that spectral distortions start to appear at 7.5 kHz on the contralateral side. The third column shows that an increase in the number of virtual loudspeakers up to $L = 1{,}962$, which would be sufficient to cover the audible frequency range, slightly improves the accuracy by smoothing the artifacts at high frequencies, specially on the ipsilateral side and the median plane. However, given that accuracy improves for one side only, there is no guarantee that the interaural accuracy also improves. The fourth column shows, on the other hand, that increasing the number of microphones



(a) Reference (b) $Q = 252$, $L = 252$ (c) $Q = 252$, $L = 1962$ (d) $Q = 1962$, $L = 252$ (e) $Q = 1962$, $L = 1962$

(f) Reference (g) $Q = 252$, $L = 252$. (h) $Q = 252$, $L = 1962$. (i) $Q = 1962$, $L = 252$. (j) $Q = 1962$, $L = 1962$.

Fig. 8 Left ear HRTFs on the horizontal (top row) and median (bottom row) planes, synthesized using Q microphones and L virtual loudspeakers.

up to $Q = 1{,}962$ has a better smoothing effect in both the ipsilateral and contralateral sides, as well as on the median plane. Nevertheless, as shown in the last column, 1962 microphones and virtual loudspeakers would be needed to faithfully maintain the monaural spectral cues. In all cases, a smooth and bounded behavior of the synthesized transfer functions along frequency and direction is observed.

4. CONCLUSION

Based on the simulation of a head model valid up to 16.6 kHz, the effects of using different numbers of microphones and virtual loudspeakers on binaural synthesis were evaluated. The placement of microphones and virtual loudspeakers was determined by following a sampling of the sphere based on the geometry of an icosahedron. Accuracy was evaluated for dense sets of sound source directions. Overall error metrics for monaural and interaural spectral cues were used.

In general terms, bounded and smooth synthesis of monaural and interaural spectral cues was possible by frequency-dependent order limitation and regularization. However, the performance of the spherical harmonic expansion significantly affected the accuracy in regions where the binaural signals showed rapid variations as a function of frequency and direction. The performance of binaural synthesis based on spherical microphone arrays was found to depend mainly on the number of microphones; this determines the maximum frequency that can be resolved by the system. Nevertheless, our results showed a limit after which increasing the number of microphones does not lead to an improvement in accuracy. Furthermore, different limits were found depending on the type of spectral cue under consideration.

The number of microphones required to synthesize the interaural cues was higher than that required to synthesize the monaural cues. The reason is that interaural synthesis requires the control of sound pressures at two ears, while monoaural synthesis requires only one ear. However, it is not simple to provide a quantitative relation between the limit to the number of microphones depending on the type of spectral cue under consideration. Such relation would depend on several parameters such as the frequency and position of the sound sources, and the distributions of the microphones. Under the particular conditions for the evaluation considered in this study, the following ordering relations have been found. When the number of virtual loudspeakers was sufficiently large to cover frequencies up to 16.6 kHz, we found that the number of microphones required to improve the overall synthesis accuracy of the interaural level difference cues was higher than the number required to improve the overall synthesis accuracy of the monaural cues. Furthermore, the number of microphones required to accurately synthesize the interaural group delay cues was greater than the number required by the interaural level difference cues.

Further considerations regarding the synthesis of individual features in the HRTFs, as well as perceptual evaluations by means of detectability of differences, and localization tests along azimuth and elevation angles, could provide more insight into the validity of the present results.

ACKNOWLEDGMENTS

The authors wish to thank Makoto Otani for his efforts in developing the BEM solver used to generate the reference HRTF data. This study was supported by a Grant-in-Aid of JSPS for Scientific Research (no. 24240016 and 16H01736), the Foresight Program for "Ultra-realistic acoustic interactive communication on next-generation Internet," and the Cooperative Research Project Program of RIEC Tohoku University (H24/A14).

REFERENCES

[1] Y. Suzuki, T. Okamoto, J. Treviño, 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," *Interdiscip. Inf. Sci.*, **18**, 71–82 (2012).

[2] M. Morimoto and Y. Ando, "On the simulation of sound localization," *J. Acoust. Soc. Jpn.* (*E*), **1**, 167–174 (1980).

[3] H. Møller, "Fundamentals of binaural technology," *Appl. Acoust.*, **36**, 171–218 (1992).

[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.* (*J*), **45**, 756–766 (1989) (in Japanese), English translation: *J. Acoust. Soc. Jpn.* (*E*), **12**, 203–216 (1991).

[5] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, revised ed. (Cambridge, MA, London, 1997) MIT Press.

[6] 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).

[7] M. Morimoto, Y. Ando and Z. Maekawa, "On head-related transfer function in distance perception," *Proc. Spring Meet. Acoust. Soc. Jpn.*, pp. 137–138 (1975) (in Japanese).

[8] M. Morimoto, N. Johren, Y. Ando and Z. Maekawa, "On head-related transfer functions," *Trans. Tech. Comm. Psychol. Physiol. Acoust.*, H-31-1-2 (1976) (in Japanese).

[9] D. S. Brungart and W. M. Rabinowitz, "Auditory localization of nearby sources. Head-related transfer functions," *J. Acoust. Soc. Am.*, **106**, 1465–1479 (1999).

[10] 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 (2005).

[11] 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**, 910–924 (2008).

[12] 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. Univers. Commun.*, pp. 429–434 (2008).

[13] E. Rasumow, M. Blau, S. Doclo, M. Hansen, S. Van de Par, D. Püschel and V. Mellert, "Least squares versus non-linear cost functions for a vitual artificial head," *Proc. Meet. Acoust.*, Vol. 19, No. 1 (2013).

[14] 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. Meet. Acoust.*, Vol. 19, No. 1 (2013).

[15] 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**, 2711–2721 (2013).

[16] B. Bernschütz, A. V. Giner, C. Pörschmann and J. Arend, "Binaural reproduction of plane waves with reduced modal order," *Acta Acust. united Ac.*, **100**, 972–983 (2014).

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

[18] N. Shabtai and B. Rafaely, "Generalized spherical array beamforming for binaural speech reproduction," *IEEE/ACM Trans. Audio Speech Lang. Process.*, **22**, 238–247 (2014).

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

[20] 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**, 516–526 (2015).

[21] 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).

[22] G. Romigh, D. Brungart, R. Stern and B. Simpson, "Efficient real spherical harmonic representation of head-related transfer functions," *IEEE J. Sel. Top. Signal Process.*, **9**, 921–930 (2015).

[23] A. Bates, Z. Khalid and R. Kennedy, "Novel sampling scheme on the sphere for head-related transfer function measurements," *IEEE/ACM Trans. Audio Speech Lang. Process.*, **23**, 1068–1081 (2015).

[24] V. A. Morozov, *Regularization Methods for Ill-posed Problems*, M. Stessin, Ed. (CRC Press, Boca Raton, 1993).

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

[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).

[27] N. A. Gumerov, A. E. O'Donovan, R. Duraiswami and D. N. Zotkin, "Computation of the head-related transfer function via the fast multipole accelerated boundary element method and its spherical harmonic representation," *J. Acoust. Soc. Am.*, **127**, 370–386 (2010).

[28] K.-S. Lee and S.-P. Lee, "A relevant distance criterion for interpolation of head-related transfer functions," *IEEE Trans. Audio Speech Lang. Process.*, **19**, 1780–1790 (2011).

[29] A. Kulkarni, S. K. Isabelle and H. S. Colburn, "Sensitivity of human subjects to head-related transfer-function phase spectra," *J. Acoust. Soc. Am.*, **105**, 2821–2840 (1999).

[30] V. C. Raykar, R. Duraiswami and B. Yegnanarayana, "Extracting the frequencies of the pinna spectral notches in measured head related impulse responses," *J. Acoust. Soc. Am.*, **118**, 364–374 (2005).