Efficient audio rendering using angular region-wise source enhancement for 360° video

Kenta Niwa[1] *Member, IEEE*, Yusuke Hioka[2] *Senior Member, IEEE*, and Hisashi Uematsu[1] *Non-Member*

***Abstract*—In virtual reality (VR), 360° video services provided through head-mounted displays (HMDs) or smartphones are widely available. Among these, some state-of-the-art devices are able to render varying auditory location of an object perceived by the user when the visual location of the object in the video moves along with the change of the user's looking direction. Nevertheless, an acoustic immersion technology that generates binaural sound to maintain a good match between the auditory and visual localization of an object in 360° video has not been studied sufficiently. This study focuses on an approach that synthesizes semi-binaural sound being composed of *virtual sources* located in each angular region and the representative head related transfer functions (HRTFs) of each angular region. To minimize the calculation cost on audio rendering and to reduce latency in downloading data from servers, the number of angular regions should be reduced while maintaining a good match between the auditory and visual localization of an object. In this paper, we investigate the minimum number of angular regions at which it is possible to maintain a good match by conducting subjective tests using a 360° video viewing system composed of virtual images and sound sources. From the subjective tests, it was confirmed that the acoustic field should be divided into more than six equi-spaced angular regions so as to achieve natural auditory localization that matches an object's location in 360° video.**

***Index Terms*—virtual reality (VR), head-mounted display (HMD), binaural synthesizing, auditory localization, angular region-wise source separation, microphone array**

I. INTRODUCTION

VIRTUAL reality (VR), which simulates presence to make one feel as if one is in remote or virtual places, has been actively studied. Advanced VR technologies may utilize different modalities to let users feel being *immersed* in a virtual environment [3] through combinations of e.g. visual [4], haptic [5], [6], gustatory/olfactory [7], and auditory [8], [9] perceptions, using multimedia contents. Of those modalities, VR technologies using visual media have been extensively studied. Many types of 360° cameras and head-mounted displays (HMDs), which include cardboard HMDs for smartphones, have been already commercialized. Showing 360° video is possible through web video streaming services and smartphone applications [10], [11] using these devices. Compared to technologies using visual media, audio media to realize better virtual reality has recently been commercialized [11], i.e. the audio is seamlessly changed with user's view in order to match the direction the user is looking.

Much research on auditory localization has been reported. Since it is known that the minimum resolution of auditory localization is approximately 1 degree in the front area of the listener provided the head of the listener is fixed [12], [13], head related transfer functions (HRTFs) with high spatial resolution (i.e. measured at many angles) have been utilized in binaural virtual source rendering. When the listener is allowed to move his/her head arbitrarily, roughly accurate localization could be achieved even if HRTFs measured at the angles deviated from the ground truth about 15 to 30 degrees [14] were utilized. Furthermore, if visual information about the sound source was provided, the listener might obtain auditory localization that matches the looking view with even more deviated HRTFs. The ventriloquist effect reported in [15] would imply this, however, to the best of our knowledge, auditory localization with an effect of visual information has not been investigated sufficiently. In particular, there have been few previous studies that investigated the minimum spatial resolution of human auditory localization when visual information is provided. Therefore, the goal of this study is (i) to investigate how much spatial resolution is needed to maintain a good match between the auditory and visual localization of an object, and (ii) to construct a 360° video viewing system that reflects the findings from these investigations.

If source signals and head related transfer functions (HRTFs) measured from a source's positions were given, it would be possible to generate binaural sounds corresponding to the direction the user is looking in. However, this would be an unrealistic assumption since HRTFs could not be measured at all possible source positions. Thus, some approximations may be needed to realize *semi*-binaural signal representation, i.e. a sound received by two ears propagating from a finite number of virtual sound sources located at regularly spaced discrete positions (angles). Ambisonics is a method for generating semi-binaural sound and has been studied actively [16]–[21]. Recently, first-order ambisonics has been used for 360° video audio rendering in some web streaming services [11]. The basic idea of ambisonics is to pick up sound sources by forming 3-dimensional spatially orthogonal directivities, which are described by spherical harmonics [22]. Users hear sounds synthesized from the recorded sound projected onto virtual loudspeakers and the HRTFs from each virtual loudspeaker through a headphone. This allows the users to feel as if they were present at the place where the HRTFs are measured [23]. However, the structure of microphone arrays

Manuscript received 24th February, 2017; accepted 26th March, 2018.
A part of this work appeared in the proceeding of ICASSP 2016 [1] and 141-st AES convention [2].
[1]: NTT Media Intelligence Laboratories, NTT Corporation, Tokyo 180-8585, Japan (e-mail: niwa.kenta@lab.ntt.co.jp, uematsu.hisashi@lab.ntt.co.jp)
[2]: Department of Mechanical Engineering, University of Auckland, 20 Symonds Street, Auckland, New Zealand (e-mail: yusuke.hioka@ieee.org)

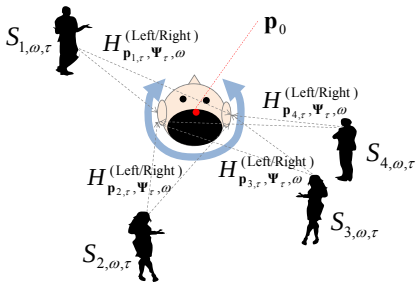


Fig. 1. Binaural signal representation using $K(=4)$ sound sources

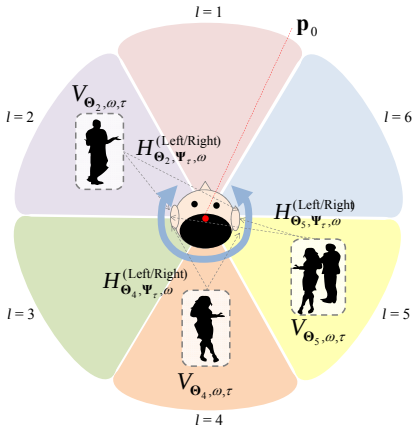


Fig. 2. Semi-binaural signal representation using angular region-grouped $L(=6)$ virtual sources

is generally restricted, e.g. four cardioid microphones placed at the apex of a regular tetrahedron facing outwards are needed for first order ambisonics or many omni-directional microphones being equally spaced and embedded in a rigid sphere are needed for higher order ambisonics [24]. It may be difficult to implement such an array of microphones in 360° camera devices because the devices are often designed by prioritizing their compactness/portability. Apart from the physical restriction, it has not been thoroughly investigated how much spatial resolution is needed for the audio of 360° viewing.

Another approach for 360° video audio rendering is multiplying HRTFs pre-measured for each angular *region* with respect to the user with source signals [1], [2], [25]. Suppose each *virtual source* is located in one of the angular regions whose representative HRTF is available; semi-binaural signals that maintain a good match between the auditory and visual localization of an object can be generated. With this approach the number of the HRTFs pre-measured will be limited by the number of angular regions. Although, this is not an exact binaural signal representation, this approach would allow for using a microphone array with an arbitrary structure. Since the source signals arriving from each angular region need to be separated, sound source separation, which also works well independently of the location and statistical properties of sound sources, needs to be applied to generate audio content.

To achieve this, the power spectral density (PSD)-estimation-in-beamspace method [26], which has proven itself as an effective sound source separation method in practical environments [27], [28], has been utilized in previous studies on 360° video audio rendering [1], [2]. Although the studies revealed that sound sources separated by the PSD-estimation-in-beamspace were able to be used for rendering sound sources at the correct angles, it is still unknown how many angular regions the field needs to be divided into in order to maintain natural auditory localization that matches the object's locations in 360° video. Although it is natural to anticipate that a higher angular resolution, i.e. higher number of angular regions, in audio recording would contribute to achieving a better match with the user's view, achieving higher angular resolution will: require a microphone array composed of many microphones; increase the calculation cost for applying a HRTF; and increase latency caused by the data transmission on streaming based content viewing services. Thus, the number of angular regions (virtual sources) should be kept at a minimum that is still able to maintain a good match between the auditory and visual localization of an object.

In this study, we investigate the effect of the number of angular regions in sound source separation and the degree of match between auditory and visual localization in a 360° video viewing system. Through subjective tests, the relationships between the number of angular regions and the degree of match between auditory localization and the direction a user is facing are investigated. We first ran a subjective listening test with pre-measured HRTFs and clean source signals in order to specify the minimum angular resolution required to achieve a good match with visual images. We also investigated the minimum angular resolution when virtual source signals are given by sound source separation using a simulated microphone array.

The rest of this paper is organized as follows. In Sec. II, a framework for 360° audio rendering using virtual sources is explained. In Sec. III, we investigate the minimum number of angular regions to divide a field without significantly degrading auditory localization through a subjective listening test. Results of similar investigations obtained when virtual sources were generated by applying the source separation algorithm to array observation signals are discussed in Sec. IV. Finally, we conclude this paper in Sec. V.

II. 360° VIDEO AUDIO RENDERING USING VIRTUAL SOURCES IN ANGULAR REGIONS

Let us assume that K sound sources are placed in a field as shown in Fig. 1. The 360° camera is fixed at a position $\mathbf{p}_0$, whereas each sound source is allowed to move arbitrarily. The relative position of the k-th source at frame time τ with respect to the camera's position $\mathbf{p}_0$ is denoted by $\mathbf{p}_{k,\tau}$. The k-th source signal in the time-frequency domain is described by $S_{k,\omega,\tau}$, where ω denotes the angular frequency. Each user wears an HMD or smartphone to view 360° video so that he/she would feel as if he/she was located at the 360° camera's position, and the direction they are facing is assumed to be captured using gyro sensors implemented in the device. The user's facing direction at τ is represented in polar coordinates by $\mathbf{\Psi}_\tau = [\Psi_\tau^{(\text{Azimuth})}, \Psi_\tau^{(\text{Elevation})}]^{\text{T}}$, where $^{\text{T}}$

denotes the transposition. When the HRTFs between the k-th source position and the outer end of the left/right auditory canals are respectively denoted by $H^{(\mathrm{Left})}_{\mathbf{p}_{k,\tau},\boldsymbol{\Psi}_\tau,\omega}, H^{(\mathrm{Right})}_{\mathbf{p}_{k,\tau},\boldsymbol{\Psi}_\tau,\omega}$, binaural signals $\mathbf{b}_{\omega,\tau} = [B^{(\mathrm{Left})}_{\omega,\tau}, B^{(\mathrm{Right})}_{\omega,\tau}]^{\mathrm{T}}$ are given by

$$B^{(\mathrm{Left})}_{\omega,\tau} = \sum_{k=1}^{K} H^{(\mathrm{Left})}_{\mathbf{p}_{k,\tau},\boldsymbol{\Psi}_\tau,\omega} S_{k,\omega,\tau}, \tag{1}$$

$$B^{(\mathrm{Right})}_{\omega,\tau} = \sum_{k=1}^{K} H^{(\mathrm{Right})}_{\mathbf{p}_{k,\tau},\boldsymbol{\Psi}_\tau,\omega} S_{k,\omega,\tau}. \tag{2}$$

Although this may represent binaural signals at a point precisely, it is required to obtain/pre-measure $S_{k,\omega,\tau}$ and $H^{(\mathrm{Left/Right})}_{\mathbf{p}_{k,\tau},\boldsymbol{\Psi}_\tau,\omega}$. Since it is almost impossible to measure the HRTF for each source position particularly when the source is continuously in motion, further simplified binaural signal representation would be needed.

The basic idea of simplifying binaural signal representation is to quantize an acoustic field into $L\ (\geq 2)$ fixed angular regions as shown in Fig. 2. By taking into account that the transfer functions do not drastically vary with a slight change of source positions, grouping source angles into L different angular regions may cause no appreciable effect on a user's auditory localization, as long as L is sufficiently large. Assuming that the acoustic field is divided into L *equi-spaced* angular regions and their centre angles are described by $\boldsymbol{\Theta}_l = [\Theta^{(\mathrm{Azimuth})}_l, \Theta^{(\mathrm{Elevation})}_l]^{\mathrm{T}}$ $(l = 1, \ldots, L)$, K sound sources are grouped into L virtual sources, which are denoted by

$$V_{\boldsymbol{\Theta}_l,\omega,\tau} = \sum_{k=\arg_k \mathbf{p}_{k,\tau}\in\boldsymbol{\Theta}_l} S_{k,\omega,\tau}. \tag{3}$$

When the HRTFs from a virtual source placed in $\boldsymbol{\Theta}_l$ to the outer end of the left/right auditory canals are denoted by $H^{(\mathrm{Left})}_{\boldsymbol{\Theta}_l,\boldsymbol{\Psi}_\tau,\omega}, H^{(\mathrm{Right})}_{\boldsymbol{\Theta}_l,\boldsymbol{\Psi}_\tau,\omega}$, semi-binaural signals are calculated by

$$B^{(\mathrm{Left})}_{\omega,\tau} \approx \sum_{l=1}^{L} H^{(\mathrm{Left})}_{\boldsymbol{\Theta}_l,\boldsymbol{\Psi}_\tau,\omega} V_{\boldsymbol{\Theta}_l,\omega,\tau}, \tag{4}$$

$$B^{(\mathrm{Right})}_{\omega,\tau} \approx \sum_{l=1}^{L} H^{(\mathrm{Right})}_{\boldsymbol{\Theta}_l,\boldsymbol{\Psi}_\tau,\omega} V_{\boldsymbol{\Theta}_l,\omega,\tau}. \tag{5}$$

In this approach, $V_{\boldsymbol{\Theta}_l,\omega,\tau}$ is estimated instead of extracting each source signal and the HRTFs at the sources' positions as in (1) and (2). To not affect the matching of audio with the direction a user is looking in, it is important to carefully select L, and investigating this is the main aim of this paper.

According to the definition of HRTF, $H^{(\mathrm{Left})}_{\boldsymbol{\Theta}_l,\boldsymbol{\Psi}_\tau,\omega}$ and $H^{(\mathrm{Right})}_{\boldsymbol{\Theta}_l,\boldsymbol{\Psi}_\tau,\omega}$ will include the characteristics of reflections/diffractions caused by the physical structure of an individual's head and torso. Although some studies [29]–[31] have shown that such individual differences would affect auditory localization, we ignore the individual differences in this study; therefore, transfer functions from a discrete angle with a constant distance from a head and torso simulator (HATS) [32] are used as an implementation of $H^{(\mathrm{Left})}_{\boldsymbol{\Theta}_l,\boldsymbol{\Psi}_\tau,\omega}, H^{(\mathrm{Right})}_{\boldsymbol{\Theta}_l,\boldsymbol{\Psi}_\tau,\omega}$. Investigating how the individuality of HRTFs affects 360° video audio rendering remains as a future work.

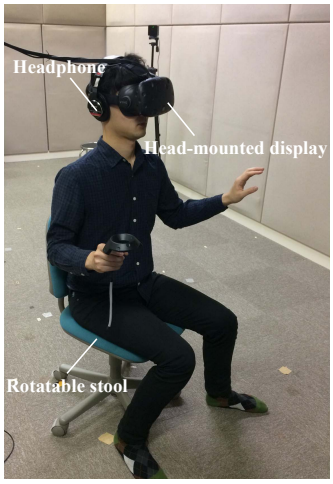


Fig. 3. Subject wearing head mount display and headphone

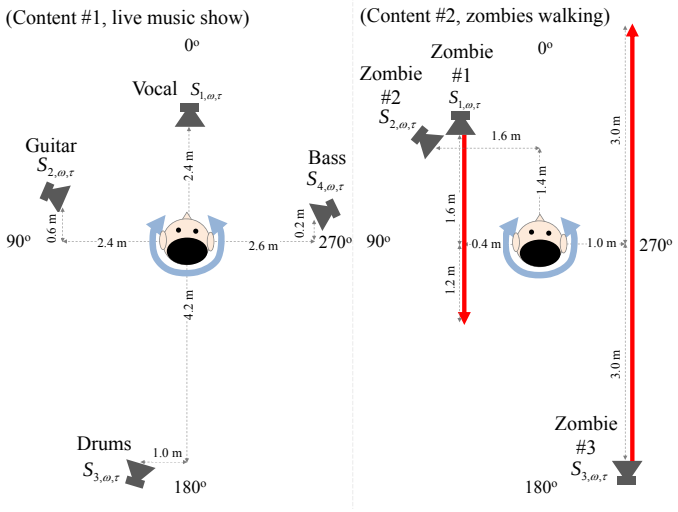


Fig. 4. Top view of audio source arrangement for two kinds of 360° video content

III. RELATIONSHIPS BETWEEN NUMBER OF ANGULAR REGIONS AND AUDITORY LOCALIZATION

A. Experimental conditions

We constructed a 360° video viewing system composed of virtual images and sound sources, the hardware configurations of which are explained below. Each subject sat on a rotatable stool and wore an HMD (HTC VIVE) and headphones (SONY MDR-CD900ST) as shown in Fig. 3. The head tracking calibration was conducted at a seated position, while the legs of the stool were in a fixed position. The HMD's resolution was 2160×1200 pixels for both eyes. To render binaural audio in real-time, an audio stream input output (ASIO) driver-based audio interface (YAMAHA UR12) was utilized. The sampling frequency of audio signals was 48 kHz. These devices were connected to a computer (CPU: Intel Core i7-6700, 3.40 GHz, GPU: NVIDIA GeForce GTX 1080, RAM: 16 GB) that had sufficient resources for rendering visual/audio smoothly. A game engine (Epic Games's Unreal Engine 4) was utilized as a visual/audio rendering platform.

Fig. 5. Examples of panorama view and user view for two kinds of 360° video content

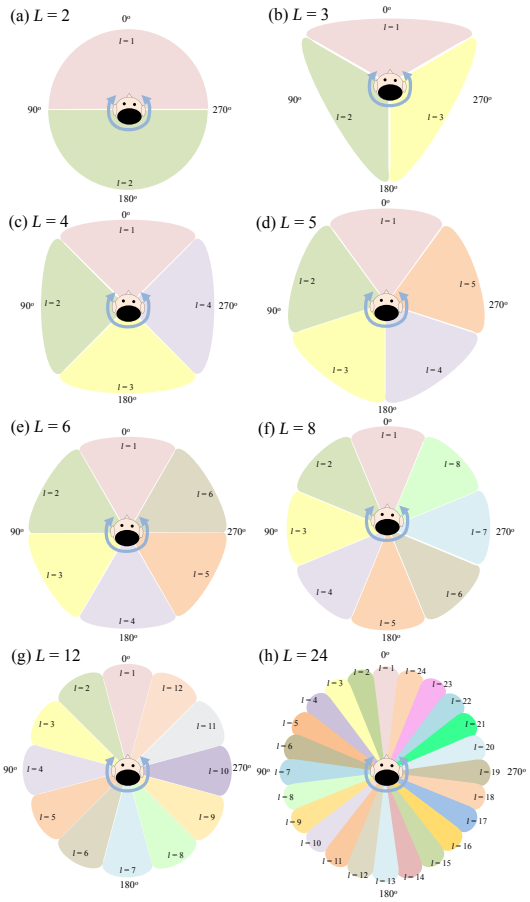


Fig. 6. Relationships between angular regions and number of angular regions L (pattern 1)

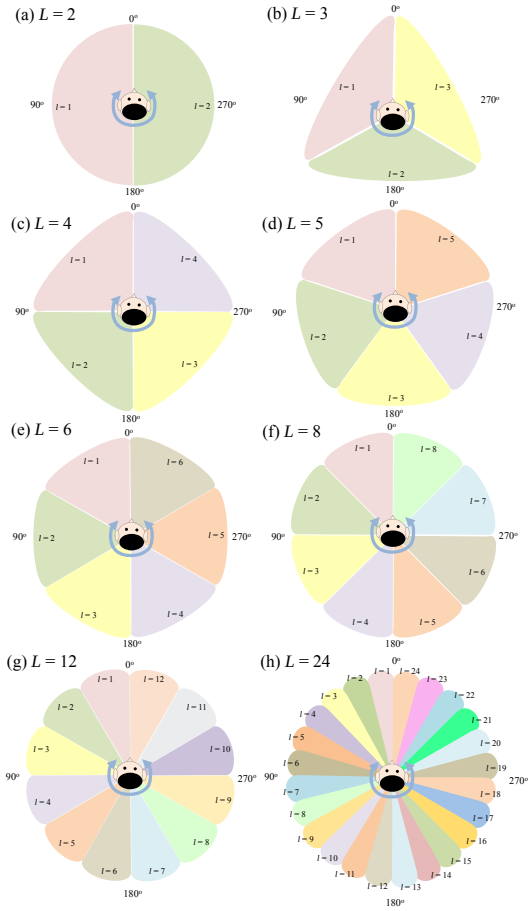


Fig. 7. Relationships between angular regions and number of angular regions L (pattern 2)

Two pieces of 360° video content were used for the experiments. *Live music show* (Content #1) was composed of four digitally generated players (sound sources), i.e. a vocalist, guitarist, drummer, and bassist, which were placed as shown in Fig. 4. A panoramic view of an entire scene as well as a view of each sound source is shown in Fig. 5. Multi-track sound sources, which were the components of a song titled "One minute smile" created by the band "Actions," were downloaded

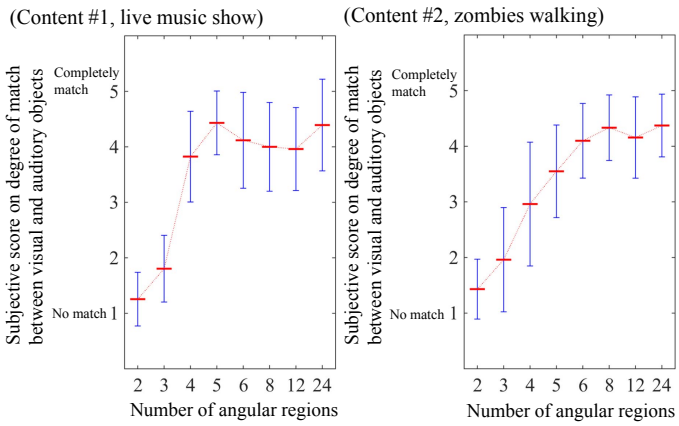


Fig. 8. Averaged subjective evaluation scores for each L when angular regions were divided into pattern 1 (Fig. 6).

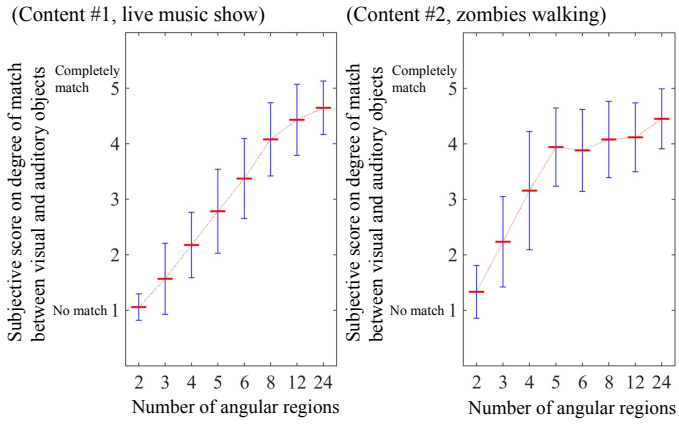


Fig. 9. Averaged subjective evaluation scores for each L when angular regions were divided into pattern 2 (Fig. 7).

from [33]. Since $S_{k,\omega,\tau}$ and $\mathbf{p}_{k,\tau}$ were known *a priori*, ideal virtual sources could be generated for each L. To make sure subjects would see movement, we made these digital factors move. Although the added motion was not particularly tuned to the sound because the motion sequence was repeated, most of the subjects responded that they did not feel any discomfort. The other piece of content was *Zombies walking in a ruined building* (Content #2), where three zombies (sound sources) are moving slowly while groaning, as shown in Fig. 4. To make it possible to discriminate between the voices of the three zombies, groaning signals were selected so that their pitches were different from each other, and this information was provided to each subject before testing. In both pieces of content, dry sound sources were used, i.e. room reverberation was not included in the virtual sources and the length of the content was 30 seconds.

There were two patterns used to divide a field into horizontal equi-spaced angular regions as shown in Figs. 6 and 7. A different number of regions was selected from $L = \{2, 3, 4, 5, 6, 8, 12, 24\}$ for each pattern. The difference between the patterns was whether there is an angular region boundary at 90 degrees or not. This difference may play an important role for auditory localization because the actual location of a sound source and the virtual location after quantizing the field into L angular regions might be drastically varied, especially when L is small.

For synthesizing semi-binaural signals, $H^{(\mathrm{Left})}_{\mathbf{\Theta}_l,\mathbf{\Psi}_\tau,\omega}$ and $H^{(\mathrm{Right})}_{\mathbf{\Theta}_l,\mathbf{\Psi}_\tau,\omega}$ were needed. To generate these transfer functions for arbitrary directions for each processing time-frame, linear interpolation in the frequency domain on the polar coordinates [34] was applied to the database [32]. The data measurement of this database was conducted by placing a HATS (B&K 4128C) in a reverberant chamber with little reverberation (T_{60}: 310 ms). The angular interval of the transfer function measurement was 5 degrees. Thus, the interpolated HRTF was smoothly changed with subject's head motion. When each sound source crosses boundary of the angular regions in Content #2, the index of virtual sources grouped together is changed. However, since the motion speed of each sound source is slow and the characteristics of the HRTFs are smoothly varied along with head motion, artifact noise did not occur in some preliminary tests conducted by the authors. Since individualized HRTFs were not used in the experiments, a preliminary study of auditory localization accuracy without video, i.e. presenting only sound stimuli generated by convolving $H^{(\mathrm{Left})}_{\mathbf{\Theta}_l,\mathbf{\Psi}_\tau,\omega}$, $H^{(\mathrm{Right})}_{\mathbf{\Theta}_l,\mathbf{\Psi}_\tau,\omega}$ and source signals, was conducted. The result, presented in Appendix, shows that auditory localization roughly matched with intended angles for many subjects, although errors caused by front-back confusion were observed, especially when the sound was propagating from front directions.

B. Subjective evaluation

Subjective tests were conducted in a semi-reverberant chamber at the NTT Musashino R&D Center, Tokyo, Japan. The background noise level and T_{60} of the chamber were 15.2 dB (A) and 180 ms, respectively. The length of the content was set to 30 seconds, which was sufficient enough for each subject to move his/her head while viewing the video. Since the sound sources are distributed horizontally, it was observed that most of the time users rotated their head horizontally rather than vertically. Each number for L was tested three times randomly for each piece of multimedia content, i.e. Live music show and Zombies walking. Seventeen subjects (5 males and 12 females, 20's - 50's) participated in evaluation tests; thus, 51 (=17 × 3) responses were obtained for each piece of content and each L. The five-point Likert scale ranging from 1 to 5 was used to evaluate the degree to which a subject's auditory localization matched the location of an object in the video (5: completely matched ↔ 1: no match). After viewing the content, each subject orally gave the subjective score to the examiner. Although each subject was instructed to face 0° every time the test started, they were allowed to arbitrarily move their head or rotate the stool in order to vary their view, i.e. they were not forced to face each object. Prior to the testing, all stimuli were presented to the subjects. During that phase, there were some subjects who searched for an object which would be effective cues to distinguish between audio signals with different L.

Figures 8 and 9 show the averaged subjective scores given when angular regions were divided into pattern 1 and 2 (Figs.

TABLE I
P-VALUE OF HYPOTHESIS TESTING SUBJECTIVE SCORES

p-values (%) of Student's paired t-test for subjective scores for each L. "*" denotes pairs of L's, the null hypothesis of which are rejected by 5% significance level after applying Bonferroni correction, i.e. individual null hypothesis, the p-value of which was smaller than $\alpha = 0.05/{}_8C_2 = 0.00179$, was rejected. Whereas, "n.s." denotes pairs of L's, which are not significantly difference each other.

Content #1, Live music show, when angular regions were divided into pattern 1 (Fig. 6)

		L						
		3	4	5	6	8	12	24
L	2	*	*	*	*	*	*	*
	3		*	*	*	*	*	*
	4			*	n.s.	n.s.	n.s.	*
	5				n.s.	n.s.	*	n.s.
	6					n.s.	n.s.	n.s.
	8						n.s.	n.s.
	12							n.s.

Content #2, Zombies walking, when angular regions were divided into pattern 1 (Fig. 6)

		L						
		3	4	5	6	8	12	24
L	2	*	*	*	*	*	*	*
	3		*	*	*	*	*	*
	4			n.s.	*	*	*	*
	5				*	*	*	*
	6					n.s.	n.s.	n.s.
	8						n.s.	n.s.
	12							n.s.

Content #1, Live music show, when angular regions were divided into pattern 2 (Fig. 7)

		L						
		3	4	5	6	8	12	24
L	2	*	*	*	*	*	*	*
	3		*	*	*	*	*	*
	4			*	*	*	*	*
	5				*	*	*	*
	6					*	*	*
	8						n.s.	*
	12							n.s.

Content #2, Zombies walking, when angular regions were divided into pattern 2 (Fig. 7)

		L						
		3	4	5	6	8	12	24
L	2	*	*	*	*	*	*	*
	3		*	*	*	*	*	*
	4			*	*	*	*	*
	5				n.s.	n.s.	n.s.	*
	6					n.s.	n.s.	*
	8						n.s.	n.s.
	12							n.s.

6 and 7), respectively. The red lines denote the averaged subjective scores for each L, and the blue error bars show their standard deviation range assuming that the scores follow a Gaussian distribution. To find the minimum number of angular regions to maintain a good match between auditory and visual localizations, statistical hypothesis tests were also conducted. Table I shows the results of a Student's paired t-test conducted for all pairs of L with a 95% confidence interval ($\alpha = 0.05$) after applying Bonferroni correction, which is a common correction method for a statistical hypothesis test that involves multiple comparisons. Thus, null hypothesis under significant level $\alpha = 0.05/{}_8C_2 = 0.00179$ was rejected.

From the test results in Figures 8 and 9, overall, it was confirmed that the subjective scores increased along with the increase of L; however, a slightly different trend can be seen between the pieces of content and the pattern of the angular regions. For Content #1, when the field was divided into pattern 1 (Fig. 6), the subjective score curve was saturated around $L = 5$, and its averaged score when $L = 5$ was 4.4. Through the statistical hypothesis tests in Table I, a significant difference in subjective score distributions could not be observed when L was greater than 5. A few exceptions where p-values were smaller than α, such as combinations of $L = \{4, 24\}$ and $L = \{5, 12\}$, can be seen, which might have occurred because L in each pair were quite different. However, when the field was divided into pattern 2 (Fig. 7), the subjective score continuously increased along with the increase of L. The causes may be that i) four sound sources were placed at fixed positions as shown in Fig. 4 and ii) the differences in the arrival direction of each sound source were drastically varied by quantizing the field into pattern 2 (Fig. 7), especially when L was small. Therefore, it may be necessary to carefully decide how to divide the angular regions when the positions of sound sources are fixed.

In comparison, with Content #2, similar subjective score curves were obtained independently of the two patterns of angular region division. Since some of the zombies moved across several angular regions, quantization errors caused by the source arrival direction may not have significantly affected the auditory localization, compared with the cases in which sound sources are placed at fixed positions. Through the statistical hypothesis tests the results of which are shown in Table I, no significant difference in subjective score distributions could be observed when L is greater than 5 given that the field is divided as pattern 1 (Fig. 6), and when L is greater than 6 given that the filed is divided as pattern 2 (Fig. 7), respectively.

Although the number of pieces of content was limited by only two, we conclude from these experimental results that i) the field should be divided into around $L \geq 6$ equi-spaced angular regions to maintain a good match between the auditory and visual localization of an object in 360° video and that ii) the way of defining the angular regions is important, especially when the positions of sound sources are stationary. These results may be consistent with the existing study [15] associated with the ventriloquist effect, which reported that roughly accurate localization could be obtained even when auditory stimulus is deviated within 30 degrees to the left and right from the ground truth. Although it would support our experimental results in some aspects, the testing environment does not perfectly match because we only used the system that can control binaural rendering corresponding to head motion.

IV. AUDIO RENDERING USING ANGULAR REGION-WISE SOURCE SEPARATION

As a method for generating virtual sources, applying angular region-wise source enhancement to array observation signals may be useful. In Sec. IV-A, a framework for generating virtual sources by applying beamforming and Wiener post-filter is explained. An implementation of Wiener post-filter

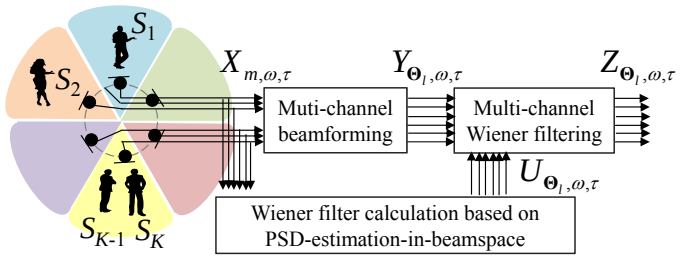


Fig. 10. Angular region-wise source separation based on beamforming with Wiener post-filtering

calculation is introduced in Sec. IV-B. In Sec. IV-C, we investigate the source separation performance and the degree of matching between auditory and visual localization, provided that some errors caused by the array signal processing are included in the virtual sources, in a subjective test similar to that presented in Sec. III.

A. Virtual sources generation using microphone array

Let us assume that an M (≥ 2)-sensor microphone array is placed near a 360° camera. When the room transfer function between the k-th sound source position and the m-th microphone is denoted by $A_{\mathbf{p}_{k,\tau},m,\omega}$, the observed signal of the m-th microphone $X_{m,\omega,\tau}$ is modeled by

$$X_{m,\omega,\tau} = \sum_{k=1}^{K} A_{\mathbf{p}_{k,\tau},m,\omega} S_{k,\omega,\tau}, \tag{6}$$

where the power of ambient background noises and/or internal sensor noise observed by the array is assumed to be negligibly small compared with that of the sound sources.

The source separation algorithm considered here is based on the framework of beamforming with Wiener post-filtering [35] as shown in Fig. 10, and the Wiener post-filter is calculated by the PSD-estimation-in-beamspace method [26]. Given that L beamformers are applied to the microphone-observed signals $X_{m,\omega,\tau}$, the output signal of the l-th ($l = 1, \ldots, L$) beamformer is given by

$$Y_{\mathbf{\Theta}_l,\omega,\tau} = \sum_{m=1}^{M} W^*_{\mathbf{\Theta}_l,m,\omega} X_{m,\omega,\tau} = \sum_{k=1}^{K}\sum_{m=1}^{M} W^*_{\mathbf{\Theta}_l,m,\omega} A_{\mathbf{p}_{k,\tau},m,\omega} S_{k,\omega,\tau}, \tag{7}$$

where * and $W_{\mathbf{\Theta}_l,m,\omega}$ denote the complex conjugate and the filter coefficient of the l-th beamformer applied to the m-th microphone, respectively. It is known that applying the Wiener post-filter to the minimum variance distortion-less response (MVDR) beamforming output [36] is the optimal way to reduce the residual noise power [37]. Therefore, the filter coefficients $\mathbf{w}_{\mathbf{\Theta}_l,\omega} = [W_{\mathbf{\Theta}_l,1,\omega}, \ldots, W_{\mathbf{\Theta}_l,M,\omega}]^{\mathrm{T}}$ are calculated by using the MVDR method, i.e. the coefficients are designed so as to minimize the beamforming output signal power while constraining the directivity gain to the angle $\mathbf{\Theta}_l$, given by

$$\mathbf{w}_{\mathbf{\Theta}_l,\omega} = \frac{\mathbf{R}^{-1}_{\omega}\mathbf{h}_{\mathbf{\Theta}_l,\omega}}{\mathbf{h}^{\mathrm{H}}_{\mathbf{\Theta}_l,\omega}\mathbf{R}^{-1}_{\omega}\mathbf{h}_{\mathbf{\Theta}_l,\omega}}, \tag{8}$$

where $\mathbf{h}_{\mathbf{\Theta}_l,\omega} = [H_{\mathbf{\Theta}_l,1,\omega}, \ldots, H_{\mathbf{\Theta}_l,M,\omega}]^{\mathrm{T}}$ denotes an array manifold vector [38], and its model under plane wave propagation is given by

$$H_{\mathbf{\Theta}_l,m,\omega} = \exp(j\omega\nu_{\mathbf{\Theta}_l,m}). \tag{9}$$

$\nu_{\mathbf{\Theta}_l,m}$ denotes the relative time delay between the m-th microphone and the array center when a plane wave is arriving from $\mathbf{\Theta}_l$. Although the spatial correlation matrix $\mathbf{R}_{\omega}$ is generally defined as the variance-covariance matrix of $\mathbf{x}_{\omega,\tau}$, it is modeled by (10) assuming that uncorrelated sound sources are arriving isotropically:

$$\mathbf{R}_{\omega} = \sum_{n=1}^{N} \mathbf{h}_{\mathbf{\Theta}_n,\omega}\mathbf{h}^{\mathrm{H}}_{\mathbf{\Theta}_n,\omega}, \tag{10}$$

where N is assumed to be a large number to model isotropical wave propagation.

The output of source separation $Z_{\mathbf{\Theta}_l,\omega,\tau}$ is then obtained by multiplying the Wiener post-filter $U_{\mathbf{\Theta}_l,\omega,\tau}$ with the MVDR beamforming outputs as

$$Z_{\mathbf{\Theta}_l,\omega,\tau} = U_{\mathbf{\Theta}_l,\omega,\tau} Y_{\mathbf{\Theta}_l,\omega,\tau}. \tag{11}$$

$Z_{\mathbf{\Theta}_l,\omega,\tau}$ ($l = 1, \ldots, L$) will be used as the l-th virtual source defined in Sec. II, i.e. $V_{\mathbf{\Theta}_l,\omega,\tau} = Z_{\mathbf{\Theta}_l,\omega,\tau}$. The following Sec. IV-B summarizes the estimation of the Wiener filter.

B. Wiener filter design using PSD-estimation-in-beamspace

Assuming that source signals are uncorrelated with each other, the PSD of the l-th beamforming output is modeled by

$$\begin{aligned}
\phi_{Y_{\mathbf{\Theta}_l},\omega} &= \left\langle |Y_{\mathbf{\Theta}_l,\omega,\tau}|^2 \right\rangle \\
&\approx \sum_{k=1}^{K}\sum_{m=1}^{M} |W^*_{\mathbf{\Theta}_l,m,\omega} A_{\mathbf{p}_{k,\tau},m,\omega}|^2 \left\langle |S_{k,\omega,\tau}|^2 \right\rangle \\
&\approx \sum_{i=1}^{L}\sum_{m=1}^{M} |W^*_{\mathbf{\Theta}_l,m,\omega} H_{\mathbf{\Theta}_i,m,\omega}|^2 \left\langle |V_{\mathbf{\Theta}_i,\omega,\tau}|^2 \right\rangle \\
&\approx \sum_{i=1}^{L} |D_{\mathbf{\Theta}_l,\mathbf{\Theta}_i,\omega}|^2 \phi_{V_{\mathbf{\Theta}_i},\omega},
\end{aligned} \tag{12}$$

where $\langle\cdot\rangle$ and $\phi_{V_{\mathbf{\Theta}_i},\omega}$ denote the expectation operation and the PSD of the i-th virtual source, respectively, and $D_{\mathbf{\Theta}_l,\mathbf{\Theta}_i,\omega} := \sum_{m=1}^{M} W^*_{\mathbf{\Theta}_l,m,\omega} H_{\mathbf{\Theta}_i,m,\omega}$ is the directivity gain of the l-th beamformer to the angle $\mathbf{\Theta}_i$. The relationship in (12) can be extended to the L beamformers as

$$\underbrace{\begin{bmatrix} \phi_{Y_{\mathbf{\Theta}_1},\omega} \\ \vdots \\ \phi_{Y_{\mathbf{\Theta}_L},\omega} \end{bmatrix}}_{\mathbf{\Phi}_{Y_\omega}} = \underbrace{\begin{bmatrix} |D_{\mathbf{\Theta}_1,\mathbf{\Theta}_1,\omega}|^2 & \cdots & |D_{\mathbf{\Theta}_1,\mathbf{\Theta}_L,\omega}|^2 \\ \vdots & \ddots & \vdots \\ |D_{\mathbf{\Theta}_L,\mathbf{\Theta}_1,\omega}|^2 & \cdots & |D_{\mathbf{\Theta}_L,\mathbf{\Theta}_L,\omega}|^2 \end{bmatrix}}_{\mathbf{D}_\omega} \underbrace{\begin{bmatrix} \phi_{V_{\mathbf{\Theta}_1},\omega} \\ \vdots \\ \phi_{V_{\mathbf{\Theta}_L},\omega} \end{bmatrix}}_{\mathbf{\Phi}_{V_\omega}}. \tag{13}$$

Since the beamforming filters are given, the directivity of the beamformers to the L directions, i.e. $\mathbf{D}_\omega$, is known *a priori*.

To estimate the PSD of each virtual source, the inverse problem of (13) needs to be solved. In the PSD-estimation-in-beamspace method [26], $\mathbf{\Phi}_{V_\omega}$ is estimated by

$$\hat{\mathbf{\Phi}}_{V_\omega} = \mathbf{D}^{-1}_{\omega}\mathbf{\Phi}_{Y_\omega}. \tag{14}$$

TABLE II
PARAMETERS IN ARRAY SIGNAL PROCESSING

Sampling frequency	48.0 kHz
Number of microphones, M	6
Radius of array	0.04 m
FFT window length	21.3 ms
Wiener filter flooring value, β	0.17 ($=10^{-15/20}$)

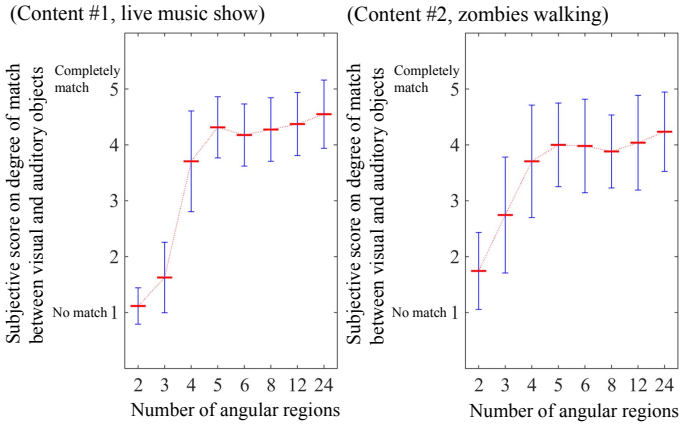


Fig. 11. Averaged subjective evaluation scores for each L when angular regions were divided into pattern 1 (Fig. 6)

For practical implementation, the source PSDs are estimated in each time-frame, assuming that source signals are temporally sparse [39]

$$\hat{\mathbf{\Phi}}_{V_{\omega,\tau}} = \mathbf{D}_{\omega}^{-1}\mathbf{\Phi}_{Y_{\omega,\tau}}. \tag{15}$$

The Wiener post-filter used to emphasize the virtual source arriving from $\mathbf{\Theta}_l$ is calculated by

$$U_{\mathbf{\Theta}_l,\omega,\tau} = \frac{\hat{\phi}_{V_{\mathbf{\Theta}_l},\omega,\tau}}{\sum_{i=1}^{L}\hat{\phi}_{V_{\mathbf{\Theta}_i},\omega,\tau}}. \tag{16}$$

For 360° video audio rendering, flooring was applied to $U_{\mathbf{\Theta}_l,\omega,\tau}$ in order to avoid generating musical noises, namely

$$U_{\mathbf{\Theta}_l,\omega,\tau} = \beta \qquad (\text{if} \quad U_{\mathbf{\Theta}_l,\omega,\tau} < \beta), \tag{17}$$

where β was empirically set to 0.17 to keep the minimum level of signal suppression to −15 dB.

C. Relationships between number of angular regions and auditory localization when array processing errors are included

The microphone array observation was numerically calculated using the content presented in Sec. III. As a simulated microphone array structure, a six-sensor regular circular array was used since three to six microphones are often utilized in practical microphone array implementations and it would be better to use a large number of microphones. To stably generate separated signals when the acoustic field is divided into five or six equi-angular regions according to the experimental results in Sec. III-B, six sensor-array was utilized. The radius of the simulated microphone array was 0.04 m. The simulated microphone array was assumed to be placed at the same position as that of the 360° camera in the VR content in Sec. III, and observation signals were calculated under the free field. Since we could not afford accurate information regarding the acoustic properties of the environment within the 3D contents, it was not possible to incorporate acoustics properties of the environment into the signal generation model. By applying angular region-wise source separation processing to those signals, virtual sources were generated for each of the L angular regions and pieces of content. The array signal processing parameters were adjusted to divide the field into pattern 1 (Fig. 6). Other parameters used in the array signal processing are listed in Table II.

1) Source separation performance: The source separation performance when generating virtual sources was measured by using the signal-to-interference noise ratio (SIR), which is listed in Table III. The power of a target source in an angular region may have become nil (there was no sound source in the angular region) when L was large, which would have made it impossible to calculate the SIR for the angular region. Thus, the power ratio of the averaged target sources in all angular regions and that of other sources was calculated for each L. The averaged SIR improvement scores (= averaged output SIR - averaged input SIR) were calculated for the overall source separation, MVDR beamforming, and Wiener post-filter, separately. In the results, the averaged SIR improvement peaked at $L = 4$ for Content #1 and $L = 5$ for Content #2, respectively. In general, the performance of the source separation algorithm would be maximized when the number of sound sources and L were equal, i.e. the separation problem is *determined*. The averaged SIR improvement was maximized around $L = 4$ since the two pieces of content included four sound sources. Nevertheless, even if the problem were over-determined ($L > 4$), the averaged SIR improvement of which would be degraded, the auditory localization would not be significantly affected because every generated virtual source would have been used to synthesize the semi-binaural signals in any case. This point will be further discussed later by looking at the results of a subjective listening test.

2) Subjective evaluation results: Subjective tests similar to that conducted in Sec. III were conducted in order to investigate the degree of matching between the auditory and visual localization of an object when array processing errors were included in the virtual sources. Experimental conditions were the same as in Sec. III-B, except the semi-binaural signals presented to the subjects were generated by using the method explained in this section. Figure 11 shows the averaged subjective scores for the localization matches for each L. Similar to the experimental results presented in Sec. III-B, the subjective scores increased along with the increase of L up to a point and then saturated. For both pieces of content, the subjective score seems to saturate around $L = 5$. From the statistical hypothesis test results shown in Table IV, a significant difference in subjective score distribution could not be observed when L is 5 or greater for content # 1 and 4 or greater for Content #2.

V. CONCLUSION

This study investigated the minimum number of angular regions a field can be divided into without significantly degrading the matching between the auditory and visual lo-

TABLE III
EVALUATION OF SOURCE SEPARATION PERFORMANCE

Content #1, Live music show, when angular regions were divided into pattern 1 (Fig. 6)

Number of angular regions, L	2	3	4	5	6	8	12	24
Averaged input SIR [dB]	0.00	-3.01	-4.76	-6.01	-6.98	-8.44	-10.40	-13.61
Averaged SIR improvement (only beamforming) [dB]	2.10	3.22	3.35	3.11	2.65	2.69	2.49	2.31
Averaged SIR improvement (Wiener filtering after beamforming) [dB]	4.87	11.06	13.28	12.17	6.26	7.70	4.48	2.35

Content #2, Zombies walking, when angular regions were divided into pattern 1 (Fig. 6)

Number of angular regions, L	2	3	4	5	6	8	12	24
Averaged input SIR [dB]	0.00	-3.01	-4.78	-6.03	-7.00	-8.46	-10.43	-13.65
Averaged SIR improvement (only beamforming) [dB]	4.37	4.49	4.07	4.17	4.16	4.34	4.08	3.95
Averaged SIR improvement (Wiener filtering after beamforming) [dB]	8.00	9.52	9.19	9.83	8.88	8.93	5.02	3.95

TABLE IV
P-VALUE OF HYPOTHESIS TESTING SUBJECTIVE SCORES

p-values (%) of Student's paired t-test for subjective scores for each L. "*" denotes pairs of L's, the null hypothesis of which are rejected by 5% significance level after applying Bonferroni correction, i.e. individual null hypothesis, the p-value of which was smaller than $\alpha = 0.05/{}_8C_2 = 0.00179$, was rejected. Whereas, "n.s." denotes pairs of L's, which are not significantly difference each other.

Content #1, Live music show, when angular regions were divided into pattern 1 (Fig. 6)

		L						
		3	4	5	6	8	12	24
L	2	*	*	*	*	*	*	*
	3		*	*	*	*	*	*
	4			*	n.s.	*	*	*
	5				n.s.	n.s.	n.s.	n.s.
	6					n.s.	n.s.	*
	8						n.s.	n.s.
	12							n.s.

Content #2, Zombies walking, when angular regions were divided into pattern 1 (Fig. 6).

		L						
		3	4	5	6	8	12	24
L	2	*	*	*	*	*	*	*
	3		*	*	*	*	*	*
	4			n.s.	n.s.	n.s.	n.s.	n.s.
	5				n.s.	n.s.	n.s.	n.s.
	6					n.s.	n.s.	n.s.
	8						n.s.	n.s.
	12							n.s.

calization of an object when viewing 360° video. Through subjective tests, it was confirmed that auditory localization that matched an object's locations in 360° video was achieved by i) dividing the field into at least around six equi-spaced angular regions and ii) carefully selecting the way of defining the angular regions, especially when the sound sources are stationary. A similar investigation was also conducted when array processing errors were included in the generation of virtual sources. When applying an angular region-wise source separation algorithm to the observation signals captured by a simulated six-sensors circular microphone array, appropriate auditory localization was obtained by dividing the field into five or more equi-spaced angular regions.

Future work may be needed to investigate the effect of individualizing HRTFs as well as including room reverberation in the HRTFs. The performance when other kinds of localization effects such as panning are utilized also has to be investigated. Further tests by increasing the numbers of pieces of content would help in finding more generic rules for specifying the optimum value for L and for determining how the division of the angular regions for the elevation angle has to be made.

ACKNOWLEDGMENTS

We thank Kouki Mitsuishi, the NTT Advanced Technology Corporation for constructing the VR system, Ken'ichi Yamanishi for constructing the evaluation system and assisting in the subjective tests, and Takanori Nishino, associate professor of Mie University, for allowing us to use the database [32].

APPENDIX
PRELIMINARY STUDY ON LOCALIZATION ACCURACY WHEN DISPLAYING ONLY SOUND STIMULUS

To verify the adequacy of using the HRTFs measured by the HATS, which generated the stimuli for the subjective listening tests, a pre-investigation of auditory localization accuracy was conducted when only sound stimuli were displayed. The stimuli were generated by multiplying the HRTFs in the database [32], and five different sound sources [1: pink noise, 2: vocalist, 3: guitarist, 4: groaning zombie (high tone), 5: groaning zombie (low tone)], the length of which was 2 seconds. In total, 17 subjects (5 males and 12 females) determined the direction of arrival (DOA) by using 12 angles from which they were told that sound was propagating from (30 degrees interval in the horizontal plane). Since 5 trials were conducted for each condition, 300 (=12×5×5) sound stimuli were displayed for each subject.

Table V shows the results of the experiment on auditory localization accuracy. The numbers in the grey-shaded areas denote the percentages of responses that matched the ground-truth. Although it seems that accurate auditory localization could be achieved only with auditory information, errors caused by front-back confusion occurred, especially when the sound was propagating from front directions (0, 30, 330 degrees). This might have happened because spatial cues for effectively identifying the DOA were not obtained when the sound sources were positioned near the median plane [40].

REFERENCES

[1] K. Niwa, Y. Koizumi, K. Kobayashi, and H. Uematsu, "Binaural sound generation corresponding to omnidirectional video view using angular region-wise source enhancement," *International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2016*, pp. 2852–2856, 2016.

TABLE V
AUDITORY LOCALIZATION ACCURACY USING HRTF DATABASE [32]

		Arrival direction [deg] (response rate [unit: %])											
		0	30	60	90	120	150	180	210	240	270	300	330
Arrival direction [deg] (ground truth)	0	42.1	10.6	3.1	1.4	3.3	10.6	25.6	1.4	0.0	0.0	0.2	1.6
	30	4.0	20.0	24.7	16.5	23.3	11.3	0.2	0.0	0.0	0.0	0.0	0.0
	60	0.5	14.1	35.1	36.7	12.2	1.2	0.0	0.0	0.0	0.2	0.0	0.0
	90	0.5	6.4	27.1	52.9	11.8	1.2	0.0	0.0	0.0	0.2	0.0	0.0
	120	0.5	2.8	11.5	24.0	48.0	13.2	0.0	0.0	0.0	0.0	0.0	0.0
	150	1.2	3.5	4.5	4.0	30.1	52.0	4.5	0.2	0.0	0.0	0.0	0.0
	180	19.1	0.5	0.0	0.0	0.5	1.9	62.1	12.0	1.2	0.2	0.7	1.9
	210	0.5	0.0	0.0	0.0	0.2	0.0	3.1	46.4	35.5	4.5	6.4	3.5
	240	0.0	0.0	0.0	0.0	0.0	0.0	0.2	9.6	44.5	28.7	13.9	3.1
	270	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.5	19.5	43.5	29.4	4.0
	300	0.2	0.2	0.2	0.0	0.0	0.0	0.5	7.1	28.0	30.1	24.0	9.6
	330	4.0	0.0	0.0	0.0	0.2	0.0	5.4	16.9	18.1	7.3	20.9	27.1

[2] K. Niwa, D. Ochi, A. Kameda, Y. Kamamoto, and T. Moriya, "Smartphone-based 360° video streaming/viewing system including acoustic immersion," *141-st Audio Engineering Society Convention*, 2016.

[3] G. C. Burdea and P. Coiffet, *Virtual reality technology (2nd ed.)*. Wiely-IEEE Press, 2003.

[4] T. Kanade, P. Rander, and P. J. Narayanan, "Virtualized reality: constructing virtual worlds from real scenes," *IEEE MultiMedia*, vol. 4, pp. 34–47, 1997.

[5] G. Robles-De-La-Torre, "The importance of the sense of touch in virtual and real environments," *IEEE MultiMedia*, vol. 13, pp. 24–30, 2006.

[6] G. C. Burdea, *Force and touch feedback for virtual reality*, 1996.

[7] T. Narumi, T. Kajinami, S. Nishizaka, T. Tanikawa, and M. Hirose, "Pseudo-gustatory display system based on cross-modal integration of vision, olfaction and gustation," *IEEE virtual reality conference 2011*, 2011.

[8] D. Begault, *3-D sound for virtual reality and multimedia*, 1994.

[9] C. A. Dimoulas, "Audiovisual spatial-audio analysis by means of sound localization and imaging: a multimedia healthcare framework in abdominal sound mapping," *IEEE Trans. on Multimedia*, vol. 18, pp. 1969–1976, 2016.

[10] D. Ochi, K.Niwa, A. Kameda, Y. Kunita, and A. Kojima, "Dive into remote events: omnidirectional video streaming with acoustic immersion," *23rd ACM Multimedia*, pp. 737–738, 2015.

[11] "Google daydream project: http://developers.google.com/vr/."

[12] R. Preibisch-Effenberger, "The human faculty of sound localization and its audiometric application to clinical diagnostics," Master's thesis, 1966.

[13] B. G. Haustein and W. Schimer, "A measuring apparatus for the investigation of the faculty of directional localization," *Hochfrequenztech. u. Elektroalustik*, vol. 79, pp. 96–101, 1970.

[14] D. Alais and D. Burr, "The ventriloquist effect results from near-optimal bimodal integration," *Current Biology*, vol. 14, pp. 257–262, 2004.

[15] F. Heller, J. Jevanesan, P. Dietrich, and J. Borchers, "Where are we?: evaluating the current rendering fidelity of mobile audio augmented reality systems," *18th international conference on human-computer interaction with mobile devices and services (MobileHCI 2016)*, pp. 278–282, 2016.

[16] D. H. Cooper and T. Shiga, "Discrete-matrix multichannel stereo," *Journal of Audio Engineering Society*, vol. 20, pp. 346–360, 1972.

[17] M. A. Gerzon, "Periphony: with-height sound reproduction," *Journal of Audio Engineering Society*, vol. 21, pp. 2–10, 1973.

[18] A. McKeag and D. McGrath, "Sound field format to binaural decoder with head tracking," *6th Australian Regional Convention of the Audio Engineering Society*, 1996.

[19] M. Noisternig, T. Musil, A. Sontacchi, and R. Holdrich, "3d binaural sound reproduction using a virtual ambisonic approach," *Symposium on Virtual Environments, Human-Computer Interfaces and Measurement Systems (VECIMS) 2003*, 2003.

[20] S. Bertet, J. Daniel, E. Parizet, L. Gros, and O. Warusfel, "Investigation of the perceived spatial resolution of higher order ambisonics sound fields: a subjective evaluation involving virtual and real 3d microphones," *30-th Audio Engineering Society Conference*, 2007.

[21] M. Gorzel, G. Kearney, H. Rice, and F. Boland, "On the perception of dynamic sound sources in ambisonic binaural renderings," *41-st Audio Engineering Society Conference*, 2011.

[22] E. G. Williams, "Fourier acoustics: sound radiation and nearfield acoustical holography," *Academic Press*, 1999.

[23] M. Noisternig, A. Sontacchi, M. Thomas, and R. Holdrich, "A 3d ambisonic based binaural sound reproduction system," *24th Audio Engineering Society (AES) Conference Multichannel Audio, The New Reality*, 2003.

[24] J. Daniel, M. Sebastien, and N. Rozenn, "Further investigations of high-order ambisonics and wavefield synthesis for holophonic sound imaging," *114-th Audio Engineering Society Convention*, 2003.

[25] K. Niwa, T. Nishino, and K. Takeda, "Encoding large array signals into a 3d sound field representation for selective listening point audio based on blind source separation," *International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2008*, pp. 181–184, 2008.

[26] Y. Hioka, K. Furuya, K. Kobayashi, K. Niwa, and Y. Haneda, "Underdetermined sound source separation using power spectrum density estimated by combination of directivity gain," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 21, pp. 1240–1250, 2013.

[27] Y. Hioka, K. Furuya, K. Kobayashi, and Y. Haneda, "Angular region-wise speech enhancement for hands-free speakerphone," *IEEE Trans. on Consumer Electronics*, vol. 58, pp. 1403–1410, 2012.

[28] K. Niwa, T. Kawase, K. Kobayashi, and Y. Hioka, "Psd estimation in beamspace using property of m-matrix," *International Workshop on Acoustic Signal Enhancement (IWAENC) 2016*, pp. 1–5, 2016.

[29] M. Morimoto and Y. Ando, "On the simulation of sound localization," *Journal of Acoustic Society of Japan*, vol. E1, pp. 167–174, 1980.

[30] J. C. Middlebrooks, "Individual differences in external-ear transfer functions reduced by scaling in frequency," *Journal of Acoustic Society of America*, vol. 106, pp. 1480–1492, 1999.

[31] K. Iida, Y. Ishii, and S. Nishioka, "Personalization of head-related transfer functions in the median plane based on the anthropometry of the listener's pinnae," *Journal of Acoustical Society of America*, vol. 136, pp. 317–333, 2014.

[32] "Nagoya university, head related transfer functions database,," in *http://www.sp.m.is.nagoya-u.ac.jp/HRTF/index.html*.

[33] "Cambridge music technology, the mixing secrets free multi-track download library, http://www.cambridge-mt.com/ms-mtk.htm."

[34] T. Nishino, M. Ikeda, K. Takeda, and F. Itakura, "Interpolating head related transfer functions," *in Proc. the seventh Western Pacific Regional Acoustics Conference (WESTPRAC VII)*, pp. 293–296, 2000.

[35] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," *in Proc. IEEE*, vol. 5, pp. 2578–2581, 1988.

[36] O. L. FrostIII, "An algorithm for linearly constrained adaptive array processing," *in Proc. IEEE*, vol. 60, pp. 926–935, 1972.

[37] K. U. Simmer, J. Bitzer, and C. Marro, *Microphone arrays: signal processing techniques and applications (1-st ed., chapter 3)*. Springer, 2001.

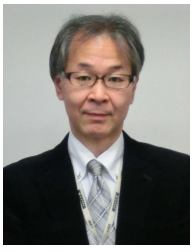
[38] D. H. Johnson and D. E. Dudgeon, *Array processing: concepts and techniques*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[39] K. Niwa, Y. Hioka, and K. Kobayashi, "Post–filter design for speech enhancement in various noisy environments," *International Workshop on Acoustic Signal Enhancement (IWAENC) 2014*, pp. 36–40, 2014.

[40] J. Blauert, *Spatial hearing the psychophysics of human sound localization (revised ed.)*. The MIT press, 1996.

Kenta Niwa (M'09) received his B.E., M.E., and Ph.D. in information science from Nagoya University in 2006, 2008, and 2014. Since joining the Nippon Telegraph and Telephone Corporation (NTT) in 2008, he has been engaged in research on microphone array signal processing as a research engineer at NTT Media Intelligence Laboratories. From 2017, he is also a visiting researcher at Victoria University of Wellington, New Zealand. He was awarded the Awaya Prize by the Acoustical Society of Japan (ASJ) in 2010. He is a member of IEEE, ASJ, and the Institute of Electronics, Information and Communication Engineers (IEICE).

Yusuke Hioka (S'04-M'05-SM'12) received his B.E., M.E., and Ph.D. degrees in engineering in 2000, 2002, and 2005 from Keio University, Yokohama, Japan. From 2005 to 2012, he was with the NTT Cyber Space Laboratories (now NTT Media Intelligence Laboratories), Nippon Telegraph and Telephone Corporation (NTT) in Tokyo. From 2010 to 2011, he was also a visiting researcher at Victoria University of Wellington, New Zealand. In 2013 he permanently moved to New Zealand and was appointed as a Lecturer at the University of Canterbury, Christchurch. Then in 2014, he joined the Department of Mechanical Engineering, the University of Auckland, Auckland, where he is currently a Senior Lecturer. His research interests include audio and acoustic signal processing especially microphone arrays, room acoustics, human auditory perception and psychoacoustics. He is a Senior Member of IEEE and a Member of the Acoustical Society of New Zealand, Acoustical Society of Japan, and the IEICE.



Hisashi Uematsu received his B.E., M.E., and Ph.D. in information science from Tohoku University, Miyagi, Japan in 1991, 1993, and 1996. Since joining the Nippon Telegraph and Telephone Corporation (NTT) in 1996, he has been engaged in research on psychoacoustics and digital sound-signal processing. He is now a senior research engineer at NTT Media Intelligence Laboratories. He was awarded the Awaya Prize by the Acoustical Society of Japan (ASJ) in 1998. He is a member of ASJ.