BINAURAL SOUND GENERATION CORRESPONDING TO OMNIDIRECTIONAL VIDEO VIEW USING ANGULAR REGION-WISE SOURCE ENHANCEMENT

K. Niwa, Y. Koizumi, K. Kobayashi, and H. Uematsu

NTT Media Intelligence Laboratories, Japan

ABSTRACT

Web applications for watching omnidirectional video through head-mounted displays (HMDs) or smartphones have been widely distributed. The goal of this study was to generate binaural sounds corresponding to the user viewpoint. Assuming that a microphone array is used for sound recording, the enhanced signal for each angular region can be extracted. By convolving head-related transfer functions (HRTFs) and enhanced signals and re-synthesizing them, binaural sounds corresponding to the user viewpoint can be virtually generated. In this paper, we propose a method for achieving angular region-wise source enhancement by generating a multichannel Wiener filter based on the *power spectral density (PSD)-estimation-in-beamspace* method. To measure user localization when watching omnidirectional video through an HMD, we used a system that enables the generation of binaural sounds corresponding to the user viewpoint in real time. Through subjective tests, we confirmed that sound localization corresponding to the user viewpoint can be obtained when applying about a 40-degree angular region-wise source enhancement.

***Index Terms*—** omnidirectional camera, microphone array, binaural sound, head-related transfer functions (HRTFs), source enhancement

1. INTRODUCTION

In virtual reality (VR), simulating presence as if being in a remote place has been actively studied. Because many approaches have been investigated to address this issue, many types of omnidirectional cameras and head-mounted displays (HMDs) have been released recently. Capturing omnidirectional video is possible through web video streaming services or smartphone applications. Despite the fact that the user viewpoint seamlessly changes, the streamed sounds through headphones are fixed even when the user viewpoint changes. Thus, the goal of this study was to generate binaural sounds corresponding to the user viewpoint.

Several methods for virtually generating binaural sounds corresponding to user control, such as those involving selective listening point (SLP) audio [1], an interactive controller [2], and instrument equalizer [3], have been studied. The SLP audio method was developed for generating binaural sounds corresponding to free viewpoint TV [4, 5]. Sounds are captured by placing many microphones to surround the acoustic field and a sound separation algorithm based on frequency-domain independent component analysis (FD-ICA) [6]-[11] is applied. After generating fixed linear filters to extract separate sources grouped into several localized regions and convolving them with head-related transfer functions (HRTFs) [12] from their positions to the listener's ears, virtually generating binaural sounds corresponding to the user control (e.g. mouse dragging) is possible. However, since FD-ICA is a batch-based algorithm, it is difficult to adaptively update filters to follow the sound source movement.

To generate binaural sounds independently of the position/number changes on sound sources, a source enhancement method for dividing the acoustic field evenly into several areas (e.g. equiangular regions) should be applied. This is preferable for application to omnidirectional video viewing since separated signals are re-synthesized after convolving with HRTFs. We previously developed the *PSD-estimation-in-beamspace* method, which may be an effective method for enhancing the signals arriving from the identified angular region [13, 14]. By using multiple beamforming output signals, the PSD of the sound sources arriving from the identified angular region and that of surrounding noise can be estimated. By designing Wiener filter using estimated PSDs, it may be possible to segregate signals arriving from the identified angular region from surrounding noise signals. Thus, by adjusting PSDs among adjacent angular regions, the multichannel Wiener filters for angular region-wise source enhancement can be generated.

For this study, we (i) expanded the PSD-estimation-in-beamspace method for angular region-wise source enhancement and (ii) evaluated our real-time omnidirectional video viewing system. When a user wears an HMD equipped with a gyro sensor, his/her head orientation can be captured in real time. By convolving the separated signals and HRTFs corresponding to the head motion and re-synthesizing them, the binaural sounds corresponding to the user viewpoint can be generated. Since the received sounds and displayed images vary with head motion, the localization of each user would adapt to the viewing content. For such a situation, we investigated the relationships between user localization with/without our angular region-wise source enhancement method through subjective tests.

This paper is organized as follows. In Sec. 2, we give a system overview of generating binaural sounds and details of our angular region-wise source enhancement method. In Sec. 3, we explain our omnidirectional video viewing system. After discussing the subjective tests we discuss our evaluation of our omnidirectional video viewing system in Sec. 4 and conclude this paper in Sec. 5.

2. BINAURAL SOUND GENERATION CORRESPONDING TO USER VIEWPOINT

2.1. System overview

Let us assume that K sound sources are observed with a microphone array composed of M sensors. The k-th source signal and m-th observation signal in the frequency ω and frame-time τ are denoted as $S_{k,\omega,\tau}$ and $X_{m,\omega,\tau}$, respectively. When the transfer function between them is denoted as $A_{m,k,\omega}$, M observation signals $\mathbf{x}_{\omega,\tau}$ are modeled as

$$\mathbf{x}_{\omega,\tau} = \mathbf{A}_{\omega}\, \mathbf{s}_{\omega,\tau} + \mathbf{n}_{\omega,\tau}, \tag{1}$$

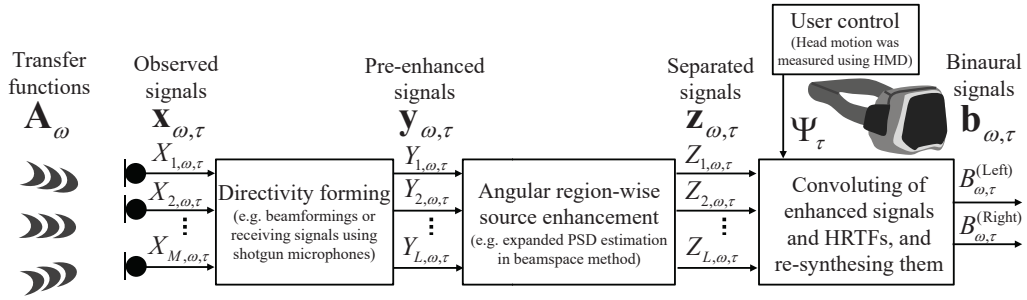

Fig. 1. Signal processing flow to generate binaural sounds corresponding to user viewpoint

where

$$\mathbf{x}_{\omega,\tau} = [X_{1,\omega,\tau}, \ldots, X_{M,\omega,\tau}]^{\mathrm{T}}, \tag{2}$$

$$\mathbf{a}_{k,\omega} = [A_{1,k,\omega}, \ldots, A_{M,k,\omega}]^{\mathrm{T}}, \tag{3}$$

$$\mathbf{A}_\omega = [\mathbf{a}_{1,\omega}, \ldots, \mathbf{a}_{K,\omega}], \tag{4}$$

$$\mathbf{s}_{\omega,\tau} = [S_{1,\omega,\tau}, \ldots, S_{K,\omega,\tau}]^{\mathrm{T}}, \tag{5}$$

$$\mathbf{n}_{\omega,\tau} = [N_{1,\omega,\tau}, \ldots, N_{M,\omega,\tau}]^{\mathrm{T}}, \tag{6}$$

Here, $^{\mathrm{T}}$ and $N_{m,\omega,\tau}$ denote the transposition and incoherent background noise at the m-th microphone, respectively.

We now explain the generation of binaural sounds corresponding to the user viewpoint, which are denoted as $\mathbf{b}_{\omega,\tau} = [B^{(\mathrm{Left})}_{\omega,\tau}, B^{(\mathrm{Right})}_{\omega,\tau}]^{\mathrm{T}}$. The user head orientation at τ is represented in polar coordinates as $\mathbf{\Psi}_\tau = [\Psi^{(\mathrm{Hor})}_\tau, \Psi^{(\mathrm{Ver})}_\tau]^{\mathrm{T}}$ and can be obtained from the gyro sensor installed in the HMD or smartphone. The directivity of the sound source and background noise are ignored. The binaural sounds corresponding to the user viewpoint are outputted by convolving HRTFs with the sound source as

$$B^{(\mathrm{Left})}_{\omega,\tau} \approx \sum_{k=1}^{K} H^{(\mathrm{Left})}_{k,\mathbf{\Psi}_\tau,\omega}\, S_{k,\omega,\tau}, \tag{7}$$

$$B^{(\mathrm{Right})}_{\omega,\tau} \approx \sum_{k=1}^{K} H^{(\mathrm{Right})}_{k,\mathbf{\Psi}_\tau,\omega}\, S_{k,\omega,\tau}, \tag{8}$$

where $H^{(\mathrm{Left})}_{k,\mathbf{\Psi}_\tau,\omega}$ and $H^{(\mathrm{Right})}_{k,\mathbf{\Psi}_\tau,\omega}$ are the HRTFs between the k-th source position and user left/right ear, respectively.

By taking into account that HRTFs do not drastically vary with slight position change, grouping source signals in a local angular region (localized signal) will have no appreciable effect on user localization. Thus, we aimed to extract L localized signals whose shaft centers are respectively toward the direction $\mathbf{\Theta}_l = [\Theta^{(\mathrm{Hor})}_l, \Theta^{(\mathrm{Ver})}_l]^{\mathrm{T}}$ $(l = 1, \ldots, L)$ instead of extracting source signal individually. Although a method of extracting localized signals, which are denoted as $Z_{\mathbf{\Theta}_l,\omega,\tau}$ $(l = 1, \ldots, L)$, is explained in Sec. 2.2, those signals are assumed to be already prepared in this section. In such a situation, the binaural sounds corresponding to the user viewpoint are approximately calculated as

$$B^{(\mathrm{Left})}_{\omega,\tau} \approx \sum_{l=1}^{L} H^{(\mathrm{Left})}_{\mathbf{\Theta}_l,\mathbf{\Psi}_\tau,\omega}\, Z_{\mathbf{\Theta}_l,\omega,\tau}, \tag{9}$$

$$B^{(\mathrm{Right})}_{\omega,\tau} \approx \sum_{l=1}^{L} H^{(\mathrm{Right})}_{\mathbf{\Theta}_l,\mathbf{\Psi}_\tau,\omega}\, Z_{\mathbf{\Theta}_l,\omega,\tau}, \tag{10}$$

where $H^{(\mathrm{Left})}_{\mathbf{\Theta}_l,\mathbf{\Psi}_\tau,\omega}$ and $H^{(\mathrm{Right})}_{\mathbf{\Theta}_l,\mathbf{\Psi}_\tau,\omega}$ are the HRTFs from the representative direction of the l-th region to the user's left/right ear, respectively. An overview of generating $\mathbf{b}_{\omega,\tau}$ is shown in Fig. 1.

Although HRTFs vary with the room reverberation time, individuality of the auricle/head structure, and distance from source to receiver [12], we ignored this for this study. This makes it possible to simply represent $H^{(\mathrm{Left})}_{\mathbf{\Theta}_l,\mathbf{\Psi}_\tau,\omega}$ and $H^{(\mathrm{Right})}_{\mathbf{\Theta}_l,\mathbf{\Psi}_\tau,\omega}$. Specifically, they are selected from a database [15] composed of impulse responses measured by discretely placing a loudspeaker and a head-and-torso simulator (HATS) in a low reverberation room.

2.2. PSD estimation in beamspace

To achieve angular region-wise source enhancement, we expanded the PSD-estimation-in-beamspace method [13]. By applying beamforming to $\mathbf{x}_{\omega,\tau}$ or simply observing signals with, e.g., shotgun microphones, the source signals arriving from $\mathbf{\Theta}_l$ are assumed to be pre-enhanced, and L pre-enhanced signals $\mathbf{y}_{\omega,\tau} = [Y_{\mathbf{\Theta}_1,\omega,\tau}, \ldots, Y_{\mathbf{\Theta}_L,\omega,\tau}]^{\mathrm{T}}$ are obtained. Assuming that the source signals are uncorrelated, the PSD of $Y_{\mathbf{\Theta}_l,\omega,\tau}$ is modeled as

$$\phi_{\mathrm{Y}_{\mathbf{\Theta}_l},\omega} = \left\langle |Y_{\mathbf{\Theta}_l,\omega,\tau}|^2 \right\rangle \approx \sum_{k=1}^{K} |D_{\mathbf{\Theta}_l,k,\omega}|^2 \phi_{\mathrm{S}_k,\omega}, \tag{11}$$

where $\langle\cdot\rangle$ and $\phi_{\mathrm{S}_k,\omega}$ denote the expectation operator and PSD of the k-th sound source, respectively. Since the relationships, expressed by Eq. (11), can be satisfied between pre-enhanced signals and localized signals, $\phi_{\mathrm{Y}_{\mathbf{\Theta}_l},\omega}$ is approximately represented as

$$\phi_{\mathrm{Y}_{\mathbf{\Theta}_l},\omega} \approx \sum_{i=1}^{L} |D_{\mathbf{\Theta}_l,\mathbf{\Theta}_i,\omega}|^2 \phi_{\mathrm{S}_{\mathbf{\Theta}_i},\omega}, \tag{12}$$

where $D_{\mathbf{\Theta}_l,\mathbf{\Theta}_i,\omega}$ denotes the average sensitivity of the l-th pre-enhancement to the angular region whose shaft center is $\mathbf{\Theta}_i$, and $\phi_{\mathrm{S}_{\mathbf{\Theta}_i},\omega}$ represents the PSD of the i-th localized signal. The relationships between $\phi_{\mathrm{S}_{\mathbf{\Theta}_i},\omega}$ and $\phi_{\mathrm{Y}_{\mathbf{\Theta}_l},\omega}$ can be modeled in matrix form as

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

To estimate L localized signals, the inverse problem of Eq. (13) can be solved. For this study, assuming that sparseness of the sound

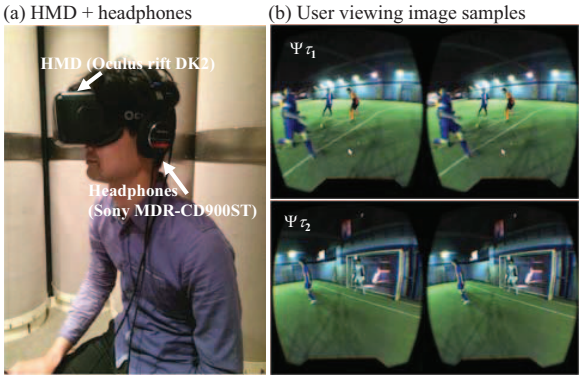

Fig. 2. (a) User watches omnidirectional video through HMD and headphones. (b) Stereoscopic view of HMD

sources in the time-frequency domain is adequately high, the PSD of the target region and that of surrounding noise can be estimated frame by frame as

$$\hat{\mathbf{\Phi}}_{\mathrm{S},\omega,\tau} = \mathbf{D}_{\omega}^{-1}\mathbf{\Phi}_{\mathrm{Y},\omega,\tau}. \tag{14}$$

Although the effect of incoherent background noise is ignored in Eq. (14) for simplicity, its PSD is separately estimated and subtracted from each PSD of the localized signal. The details of how to calculate the PSD of incoherent background noise is described in our previous work [14]. In this study, we estimated the PSD of background noise assuming that it is temporally stationary.

2.3. Angular region-wise source enhancement by adjusting estimated PSDs of localized signals

To expand the PSD-estimation-in-beamspace method for angular region-wise sound enhancement, we adjusted multichannel PSDs of localized signals between adjacent angular regions. When simply applying Eq. (14), the estimated PSDs may include estimation errors since the same sound source can be emphasized among adjacent angular regions. This may degrade the sound localization of binaural signals.

To reduce the PSD estimation error among adjacent angular regions, the region index whose estimated PSD is the highest, described by ξ_τ, is calculated for each frame as

$$\xi_\tau = \arg\max_{l} \sum_{\omega\in\Omega} \hat{\phi}_{\mathrm{S}_{\mathbf{\Theta}_l,\omega,\tau}}, \tag{15}$$

By suppressing $\hat{\phi}_{\mathrm{S}_{\mathbf{\Theta}_l,\omega,\tau}}$, whose region index is adjacent to Θ_{ξ_τ}, emphasizing the same sound source in different regions can be avoided by using

$$\hat{\phi}_{\mathrm{S}_{\mathbf{\Theta}_l,\omega,\tau}} = \begin{cases} \hat{\phi}_{\mathrm{S}_{\mathbf{\Theta}_l,\omega,\tau}} & (l = \xi_\tau) \\ 0 & (l \in \phi_\tau) \end{cases}, \tag{16}$$

where ϕ_τ denotes the index set of the adjacent angular region of Θ_{ξ_τ}. Finally, by using adjusted PSDs, a Wiener filter to enhance signals arriving from the l-th angular region is generated using

$$G_{\mathbf{\Theta}_l,\omega,\tau} = \frac{\hat{\phi}_{\mathrm{S}_{\mathbf{\Theta}_l,\omega,\tau}}}{\sum_{i=1}^{L} \hat{\phi}_{\mathrm{S}_{\mathbf{\Theta}_i,\omega,\tau}}}. \tag{17}$$

The localized signals are calculated as

Table 1. Conditions for HRTFs measurement [15]

Sampling frequency	44.1 kHz (Up-sampling to 48 kHz)
Room reverberation time	310 ms
# of horizontal angles	72 (5° interval)
# of vertical angles	28 (5° interval) (from -45 to 90 degrees)

$$Z_{\mathbf{\Theta}_l,\omega,\tau} = G_{\mathbf{\Theta}_l,\omega,\tau} Y_{\mathbf{\Theta}_l,\omega,\tau}. \tag{18}$$

By applying the inverse fast Fourier transform (FFT) to $Z_{\mathbf{\Theta}_l,\omega,\tau}$, the time-domain localized signals are obtained.

3. REAL-TIME OMNIDIRECTIONAL VIDEO VIEWING SYSTEM

The real-time omnidirectional video viewing system, which includes an HMD and headphones, as shown in Fig. 2, was used [18]. The Oculus Rift DK2 (Oculus VR Inc.) was used as the HMD. With Oculus Rift DK2, wide perspective fish-eye lenses, whose view angle is from 100 to 110 degrees, are mounted in front of the screen with a graphical resolution of 1920×1080 pixels. To compensate for the optical distortion originating from the fish-eye lenses, the inverse characteristics of the lenses are multiplied to the screen-displayed images for each frame. Also, $\mathbf{\Psi}_\tau$ can be observable in real time using the gyro sensor and triaxial sensor installed in Oculus Rift DK2. Since the image to cover the visual field is generated and corresponds to the user ' s head motion, immersive video viewing is possible for each user.

To solve the problem of large-capacity image-data transmitting and rendering, we used the H. 264 (MPEG-4 AVC)-based omnidirectional image streaming system [16, 17, 18]. The high-resolution image captured with an omnidirectional camera is too large to render to correspond to $\mathbf{\Psi}_\tau$. Instead of streaming uncompressed omnidirectional images, both the entire low-resolution omnidirectional image (around 0.5 Mbps) and partial high-resolution image corresponding to the user overlooking region (around 2.0 Mbps) are transmitted in parallel. To follow quick head motion, low-resolution images are outputted to the screen after modifying the optical distortion of the fish-eye lenses. When the user ' s head is fixed to an identified direction for around 5.0 seconds, the resolution of the user viewpoint is switched from low to high.

For outputting calculated binaural sounds, the ASIO audio interface (Roland Octa-capture) and headphones (SONY MDR-CD900ST) are used. The L localized signals are assumed calculated beforehand. By convolving HRTFs and localized signals, as in Eq. (10), binaural sounds corresponding to the user viewpoint are generated in real time. The impulse responses measured by placing HATS (B&K 4128C) in a soundproof chamber were used as HRTFs [15]. Since the angular interval of the HRTF database is 5.0 degrees, as shown in Table 1, seamless localization control corresponding to a user ' s head motion can be achieved.

4. EXPERIMENTS

4.1. Recording setup

Two kinds of futsal (a modified form of soccer) actions were recorded using an omnidirectional camera (Point Grey Ladybug3) and an icosahedral microphone array, as shown in Fig. 3. The icosahedral microphone array was composed of $M = 20$ shotgun microphones positioned for each face center. The diameter of the

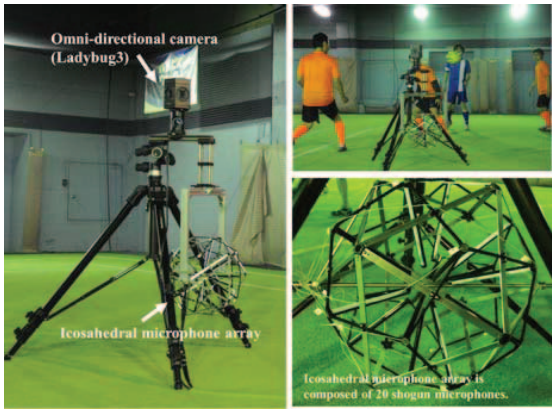

Fig. 3. Recording system composed of omnidirectional camera and icosahedral microphone array

Table 2. Experimental parameters

Sampling frequency	48.0 kHz
Number of microphones, M	20 (shotgun microphone)
Diameter of array	0.4 m
Number of beamspaces, L	20
FFT window length	5.30 ms
FFT shift	2.65 ms

array was 0.4 m. The camera and microphone array were placed on the futsal court. To prevent the array from being caught on camera, it was placed 0.5 m below the camera.

The recorded actions were (i) a practice match: players chasing a ball to score a goal and (ii) ball juggling: four players juggle a ball alternately while playing word chain. Ball kicking, running around the field, players shouting, audience cheering, and air conditioning noise were included in the observed signals. Since the motion of futsal players was fast during the practice match, the arrival direction of sounds varied from frame to frame. Since the player position was nearly fixed for ball juggling, the viewer was able to associate the arrival direction of sounds while viewing the images.

By dividing the acoustic field into $L = 20$ equiangular regions whose center shaft extends from the array center through each microphone, localized sounds were estimated. The angular difference between adjacent regions was 41.8 degrees. Since shotgun microphones were used in this experiments, the pre-enhanced signals $\mathbf{y}_{\omega,\tau}$ were equal to $\mathbf{x}_{\omega,\tau}$. Since non-stationary sound sources, such as ball kicking, were mixed with the observed signals, they were analyzed in a short window length (5.3 ms). The adjacent regions defined in Eq. (16) were determined as a hemisphere; thus, the number of simultaneous localized sounds was up to 2. The other experimental parameters are listed in Table 2.

4.2. Subjective evaluations

We conducted subjective evaluations using our omnidirectional video viewing system. We found that applying our angular region-wise source enhancement method affected (i) sound localization corresponding to the overlooking image and (ii) overall sound quality, which includes the degradation caused from our multichannel Wiener filtering. As a comparison method, pre-enhanced $M = 20$ observed signals were simply convolved with HRTFs corresponding to the user ' s head motion. The participants were six males and the sequence number for each action and the sound processing method

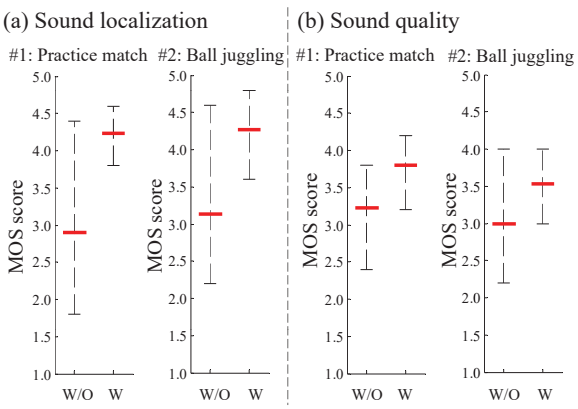

Fig. 4. Subjective evaluation results, (a) sound localization, and (b) sound quality. (W/O): Without source enhancement processing, (W): With our angular region-wise source enhancement method.

was five. The length of actions was limited to 20.0 seconds and we asked each user to rotate his head. The mean opinion scores (MOS) (5: Excellent, 4: Good, 3: Fair, 2: Poor, 1: Bad) were used as an evaluation measure.

Fig. 4 shows the MOS scores. The bold horizontal lines indicate the average scores of the six participants and the horizontal narrow lines indicate the minimum/maximum score. As shown in Fig. 4(a), the MOS score of sound localization increased by applying our angular region-wise source enhancement method. However, some participants rated high score to the comparison method. This was due to the fact that effective clues for sound localization were included in the observation signals since many directivity microphones ($M = 20$) were used for observation. As shown in Fig. 4(b), the sound quality improved a bit with our method. This was due to the fact that stationary air conditioning noise was suppressed with our method.

5. CONCLUSION

We proposed an angular region-wise source enhancement method for our immersive omnidirectional video viewing system. To generate binaural sounds corresponding to the user viewpoint, we expanded the PSD-estimation-in-beamspace method as an angular region-wise source enhancement method. By convolving enhanced localized signals and HRTFs corresponding to a user ' s head motion, binaural sounds were generated in real time. Through subjective tests using our omnidirectional video viewing system, we confirmed that sound localization corresponding to the user viewpoint could be obtained.

Some issues remain for future work such as investigation of (i) the relationships between sound localization and angular width to be separated and (ii) the effect of HRTF individuality.

6. ACKNOWLEDGEMENTS

We would like to express our gratitude to Shinnosuke Iwaki (DWANGO Co. Ltd.) for providing the omnidirectional image viewer application. We specially thank Kazuya Takeda (Nagoya University) and Takanori Nishino (Mie University) for allowing us to use the HRTF database. We received generous support from Daisuke Ochi, Akio Kameda, and Yutaka Kunita (NTT Media Intelligence Laboratories) for video coding and content recording.

7. REFERENCES

[1] 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", *in Proc. ICASSP2008*, pp.181–184, 2008.

[2] N. Kamado, H. Nawata, H. Saruwatari, K. Shikano, and T. Nomura, "Interactive controller for audio object localization based on spatial representative vector operation", *in Proc. IWAENC2010*, 2010.

[3] K. Itoyama, M. Goto, K. Komatani, T. Ogata, and H. Okuno, "Instrument equalizer for query-by-example retrieval: improving sound source separation based on integrated harmonic and inharmonic models", *in Proc. ISMIR 2008*, pp. 133–138, 2008.

[4] T. Fujii and M. Tanimoto, "Free-viewpoint TV system based on the ray–space representation", *in Proc. SPIE ITCom*, vol. 4864–22, pp. 175–189, 2002.

[5] "ISO/IEC JTC1/SC29/WG11 (N9168)," July 2007(Lausanne, Switzerland).

[6] A. Bell and T. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *in Proc. Neural Computation*, vol. 7, pp. 1129–1159, 1995.

[7] S. Choi, S. Amari, A. Cichocki, and R. Liu, "Natural gradient learning with a nonholonomic constraint for blind deconvolution of multiple channels," *in Proc. International workshop on ICA and BSS*, pp. 371–376, 1999.

[8] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *in Proc. Neurocomputing*, vol. 22, pp. 21–34, 1998.

[9] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, and K. Shikano, "Blind source separation combining independent component analysis and beamforming," *EURASIP J.Applied Sig. Proc.*, pp. 1135–1146, 2003.

[10] F. Asano, S. Ikeda, M. Ogawa, H. Asoh, and N. Kitawaki, "Combined approach of array processing and independent component analysis for blind separation of acoustic signals," *IEEE Trans. on Speech and Audio Proc.*, vol. 11, no. 3, pp. 204–215, 2003.

[11] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Solving the permutation problem of frequency-domain BSS when spatial aliasing occurs with wide sensor spacing," *in Proc. ICASSP 2006*, vol. V, pp. 77–80, 2006.

[12] J. Blauert, "Spatial hearing (revised ed.)," 1996.

[13] 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.

[14] K. Niwa, Y. Hioka, and K. Kobayashi, "Post–filter design for speech enhancement in various noisy environments," *in Proc. IWAENC 2014*, pp. 36–40, 2014.

[15] Nagoya University, Head Related Transfer Functions Database, http://www.sp.m.is.nagoya-u.ac.jp/HRTF

[16] D. Ochi, S. Iwaki, Y. Kunita, J. Hirose, K. Fujii, and A. Kojima, "HMD viewing spherical streaming system," *in Proc. ACM MM 2014*, 2014.

[17] D. Ochi, S. Iwaki, A. Kameda, Y. Kunita, and A. Kojima, "Live streaming system for omnidirectional video," *in Proc. IEEE VR 2015*, 2015.

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