Audio Engineering Society

Convention Paper 10362

Presented at the 148th Convention
Online, 2020 June 2-5

Are full-range loudspeakers necessary for the top layer of Three-Dimensional audio?

Toru Kamekawa[1] and Atsushi Marui[1]

[1]*Tokyo University of the Arts*

Correspondence should be addressed to Toru Kamekawa (kamekawa@ms.geidai.ac.jp)

ABSTRACT

When a human perceives a space by hearing, horizontal sound image localization and spaciousness are sensed based on ILD (Inter-aural level difference) and ITD (Inter-aural time difference). However, in the vertical direction, spectral cue and directional band caused by the difference in the frequency characteristic of the direction of arrival of sound caused by the shape of the ear, are crucial. The authors investigated the difference between the original 22.2 multichannel sound and its filtered sound by limiting the playback frequency band of its top layer using various contents. The results demonstrated that there were no significant differences in spatial impression, even if the top layer does not have a band below approximately 400 Hz.

1 Introduction

1.1 Purpose and background of the study

Three-dimensional (3D) audio has playback channels along its height in addition to its front, rear, left, and right. It can express spatial sound that could not be achieved using conventional two-channel stereo or 5.1 surround sound [1]. Several 3D audio formats have been standardized within ITU-R BS.2159 [2], such as NHK 22.2 channel sound, Dolby Atmos, and Auro 3D. It is recommended to use speakers having the same characteristics for reproducing such 3D audio [3]. However, it is not easy to prepare an ideal reproduction environment in a consumer's home audio, for example, providing the same quality speakers for all channels as would exist in the professional studios. This study aims to investigate how ideal playback schemes can be scaled down to such an actual playback environment.

1.2 Sense of spaciousness for 3D audio

When a human perceives a space by hearing, horizontal sound image localization and spaciousness are sensed based on the level difference and time difference between the left and right ears (ILD and ITD) [4, 5]. However, in the vertical direction, the difference in the frequency characteristic of the direction of arrival of sound caused by the shape of the ear (so-called "spectral cue" and "directional band") becomes a clue of localization [6]. Therefore, unlike the horizontal direction, the spatial perception in the vertical direction depends on the frequency spectra at the ears.

Hebrank *et al.* showed that the frequency bands required for sound image localization in the median plane was from 4k to 16 kHz [7]. Morimoto *et al.* also concluded that the bands below 4 kHz did not contribute to the median localization [8]. The authors' research also showed that a frequency band of 4 kHz or higher is necessary to perceive the vertical direction correctly [9]. Thus, the frequency band contributing to the vertical localization is a high-frequency band. The middle and low frequency bands of 4 kHz or lower are considered to contribute weekly for vertical localization.

However, regarding the sense of spaciousness, apparent source width (ASW), and listener envelopment (LEV) are well known to express this perception [10]. These horizontal spatial factors are considered to be relevant to the decorrelation between the left and the right signals. In contrast, the sense of spaciousness in the vertical direction is not related to the decorrelation of the upper and lower signals [11, 12, 13].

In studies regarding concert halls, vertical listener envelopment (VLEV) and overhead sound image perception (OSI) were defined as vertical spatial impressions [14]. Further, they were distinguished from horizontal listener envelopment (HLEV). However, it is still unclear precisely which physical quantities correspond to them. And it is also unclear what frequency band is the key to the sense of spaciousness in the vertical direction.

Therefore, in this study, we focused on the frequency band of reproduction in the vertical direction and examined the frequency limitation where the difference from the original signal could not be distinguished.

2 Experiment

In the experiment, we examined to limit the frequency band of the top layer of 22.2 multichannel sound to find out what kind of difference was felt and where the threshold of the frequency from which the difference could be perceived was.

2.1 Methods

Two experiments were conducted using either a low-cut filter or a high-cut filter for the upper nine channels of the 22.2 multichannel audio. These filters were obtained by the filtergraph function of Cycling '74 Max 8 (gain = 1.0, Q = 0.7). In experiment 1, the original sound (A) and the filtered sound (B) with the cutoff frequency of the low-cut filter at the highest frequency (15.7 kHz) were compared. The participant then wrote down what they felt about the difference between A and B on a piece of paper. Subsequently, the participant moved the slider up and down in a GUI using an iPad created in Max 8 to adjust the cutoff frequency of the filter (Fig.1). The participant then stopped the slider where the difference between A and B could not be distinguished. The position was recorded as the threshold value of the low-cut filter. The sound of the band cut in the top layer was mixed with the corresponding middle layer channel such that there was not much change in volume at the listening position (Overhead TpC which does not correspond to middle layer was assigned to the back center (BC)). Similarly, in experiment 2, the high-cut filter was set to the lowest frequency (62.5 Hz), and the participant first wrote the difference between A and B. Next, they moved the slider and answered the threshold value of the high-cut filter. Experiments 1 and 2 were conducted in random order by participants.

The frequency moved by the slider was a logarithmic scale divided into 256 steps. The resolution of the step corresponds to approximately 1/36 octave (33 cents). The range of the cutoff frequency and frequency step used for the adjustment of the low-cut and the high-cut are as follows:

- Low-cut filter
  - frequency range: 125 Hz ~15.7 kHz
  - minimum frequency step: 2.39 Hz@125 Hz
  - max frequency step: 295 Hz@15.7 kHz
- High-cut filter
  - frequency range: 62.5 Hz ~13 kHz
  - minimum frequency step: 1.32 Hz@62.5 Hz
  - max frequency step: 268 Hz@13 kHz

2.2 Stimuli of the experiment

In the listening experiment, the following eight types of stimuli were used to compare as many different contents as possible. The initials at the beginning are used below as respective abbreviations.

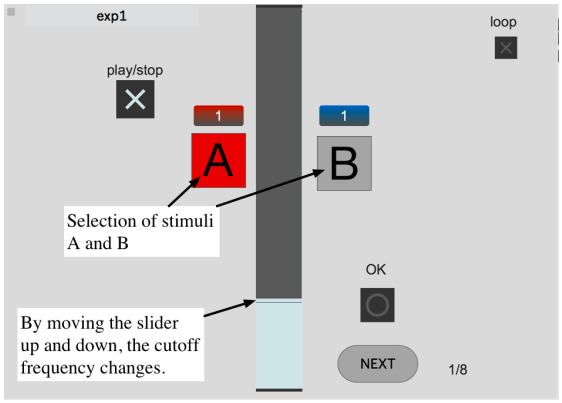

Fig. 1: GUI for the experiment using iPad

- VL: Violin solo recorded in the studio (*Gavotte en Rondeau* from Violin Partita No.3 in E major, BWV 1006 by J.S.Bach) using the spaced array technique corresponding to 22 loudspeakers' layout [15].
- ML: Orchestra recorded in the concert hall (Symphony No.5 by G. Mahler). An A-format Ambisonics microphone (SPS200) was located in the middle between the left and right main microphones. The ambisonics signal converted to 22 channels are mixed to a four-channel main microphone (FL, FR, BL, and BR) and spot microphones. The top layer is the only signal by Ambisonics.
- KD: Chorus with an orchestra recorded in the concert hall (*Kaido-tosei* by K.Nobutoki). The chorus microphones are located in the top layer
- ET: Japanese traditional music recorded in the studio (*Etenraku*). A *sho* (Japanese free-reed musical instrument), a Japanese flute, and Japanese drums are located in the top layer.
- LN: Musical work created for 22 channels with overdubbing of voice (*Lenna* by M.Hosoi). Various voices are assigned to the top layer.
- MK: A piece comprising music and poetry readings and sound effects for 22.2 multichannel audio (*My Kingdom*). Various instruments and sound effects are assigned to the upper layer.

Table 1: LAeq of original material and filtered material processed to the maximum (The unit is dB)

Stimulus	Original	low-cut	high-cut
VL	70.8	71.2	71.3
ML	76.0	76.3	76.2
KD	78.0	77.8	77.9
ET	75.0	73.9	73.8
LN	79.7	79.8	79.9
MK	76.4	76.1	76.2
BN	62.5	61.7	61.4
PN	65.2	64.7	64.7

- PN: Pink noise. Twenty-two channels are randomly generated, and the correlation between channels is low.
- BN: Burst signal of pink noise. Same as the pink noise above (PN), and it is played back at 500 ms duration and 500 ms interval.

These sources were excerpted in approximately 30 s and repeatedly played at a 96 kHz sampling rate and 24 bits resolution. LFEs were used for KD, ET, and MK.

2.3 Apparatus

The experiment was conducted using the 22.2 multichannel sound system of Studio B of Tokyo University of the Arts (floor area= 68m^2, ceiling height= 5m, reverb time= *ca*.0.4 s at 500 Hz). Figure 2 indicates the layout of loudspeakers and listening positions. Twenty-four loudspeakers were placed in the studio via KS digital C5-Coax and ADM-B2, and they corresponded to the ITU-R BS. 2051 [3]. The high-cut frequency of LEFs was 120 Hz. Figure 3 shows frequency response of the front center (FC) loudspeaker measured at the center listening position.

The playback level was set from approximately 62 dB (BN) to 80 dB (LN) LAeq at the center listening position. The difference in the sound pressure level between those filtered by each stimulus was within approximately 1 dB (Table 1).

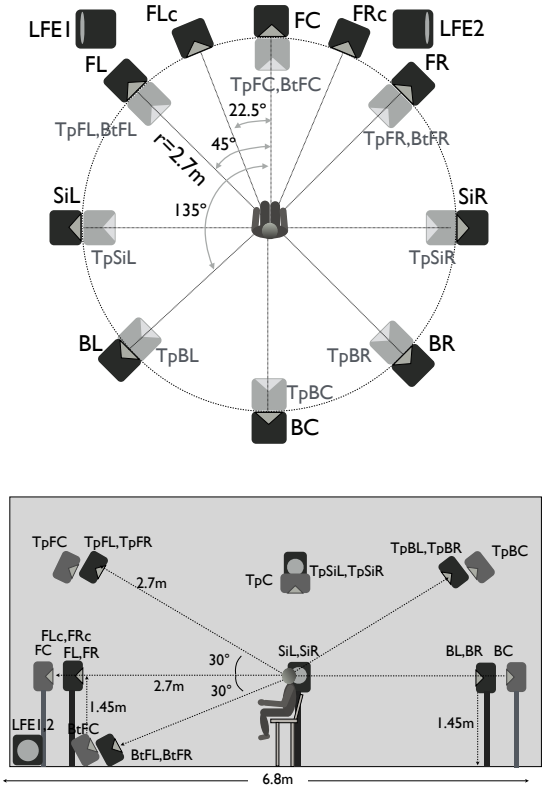

Fig. 2: Layout of the listening experiment. The upper panel indicates the horizontal direction and the lower panel indicates the section. Delays in the distance difference between BtFL, BtFC, BtFR, and TpC, were inserted to adjust the arrival time corresponding to the distances (2.7 m) of each loudspeaker, which is equal.

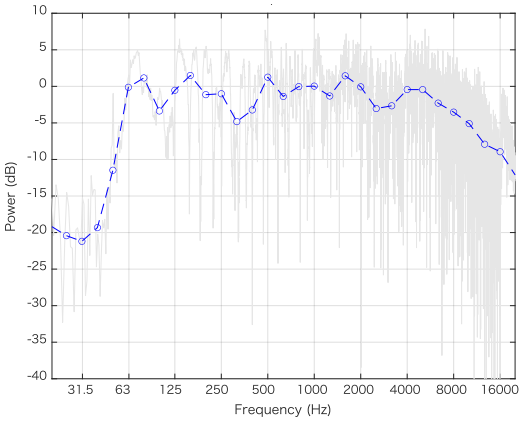

Fig. 3: Frequency response of the front center loudspeaker (FC) measured at the center listening position. The blue line indicates frequency response of 1/3 octave band.

2.4 Participants

Sixteen students and faculty staff (mean=27.9 years, sd=11.4) from our university participated in both experiments. All of them reported normal hearing and had one year or greater experience of technical ear training. The participants were allowed to listen and compare each pair as many times as desired. Martens *et al.* reported that the perception of the presence of 'height channels' requires head rolling[16]. Therefore, there was no restriction on moving one's head during listening. Each participant conducted the experiments, and a random combination was compared for each trial. Each participant conducted the experiments, and a random combination was compared for each trial.

3 Results

3.1 Low cut filter (Experiment 1)

When the low frequencies of the top layer were cut, the following comments were obtained compared to the original sound source. Abbreviations in parentheses indicate stimuli (see section 2.1).

- The top layer has become thinner (BN).
- The reverberation is gone (VL).

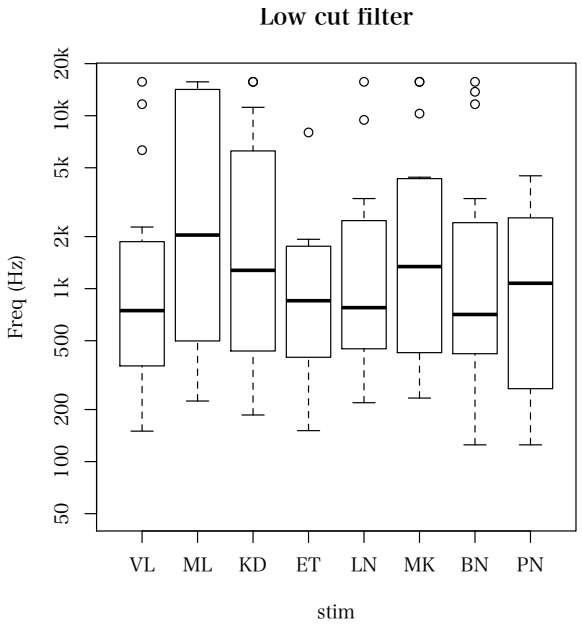

Fig. 4: Boxplot of the low cutoff frequency of the top layer for each stimulus of all participants

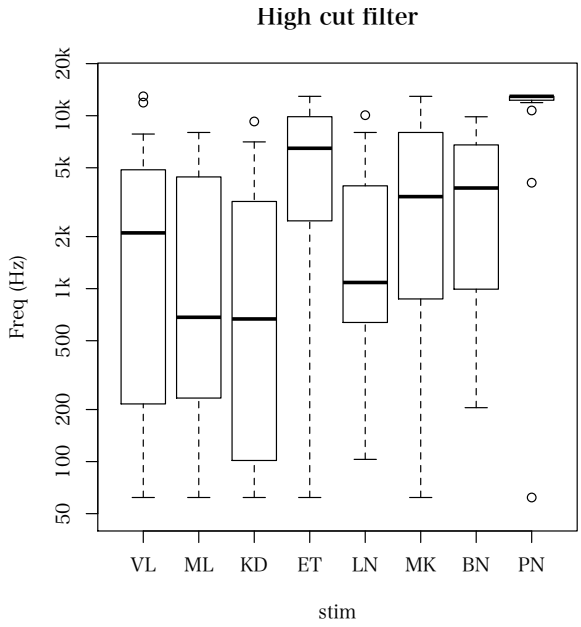

Fig. 5: Boxplot of the high cutoff frequency of the top layer for each stimulus of all participants

- The localization of the instruments is slightly different (ET, LN).
- The feeling of spaciousness disappears (VL, ML, MK).
- The sound image feels far away (KD).
- High frequencies become stronger (BN).
- The middle range sounds emphasized (MK, PN)

Figure 4 shows the result of experiment 1. The horizontal axis shows stimuli, and the vertical axis shows the threshold frequency at which differences cannot be distinguished. Since the lower side of the box in the box plot represents the first quartile, depending on the material, if there is a band above 400 Hz, 75% of listeners cannot distinguish the difference from the original.

3.2 High-cut filter (Experiment 2)

Similarly, when the high frequencies of the top layer were cut, the following comments were obtained compared to the original sound source.

- The sound seems to be stuffed (BN, ET).
- The sound image becomes narrow (ML, PN).
- The middle range is emphasized (VL, MK).
- The sound source feels far away (KD).
- The sound of the flute is a bit difficult to hear (ET).

Figure 5 shows the result of experiment 2. According to the box plot of the figure, the permissible high-cut frequency differs depending on the sound material. For pink noise, listeners notice the difference with even a slight cut.

4 Discussion

4.1 Low frequency required for the top layer

From the experimental results, the frequency band required for the top layer in the low range depends on the source. In particular, the median value of ML obtained by converting Ambisonics to 22 channels has a

higher frequency than other sources. In these cases, the reflection sound recorded in Ambisonics occupied the top layer; there is no effect of reproduction from the top layer. VL (Violin Solo) feels the effect of the top layer than ML, though only the reflection sound is assigned to the top layer like ML. The reason may be because the top layer of VL is recorded in the spaced array technique.

On the other hand, ET with the Japanese flutes localized in the top layer and LN with the voice localized in the top layer have relatively slight variation in answers and are approximately 500 Hz. The effect of the top layer is useful when the actual sound is reproduced from the top layer.

For the frequency band requirement for the top layer in the low range, 75% of participants at 400 Hz or higher, and almost everyone at 150 Hz or higher, did not distinguish the difference between the original and the filtered sound used in the experiments.

4.2 High frequency required for the top layer

Regarding the high-frequency limitation, the difference may not be noticeable depending on the content. The stimuli recorded in a hall and a studio such as VL, ML, and KD were assigned reverb and mainly ambiance. It was difficult to distinguish between the presence and the absence of the top layer. As in the low-frequency range, it is easy to notice the change in the high frequency range when the ET with the Japanese flutes is localized in the top layer. Moreover, PN, including many frequency bands, all listeners noticed a slight high-frequency change. From the comments describing the difference in presence or absence of high frequency, it is concluded that they are responsible for the difference in timbre. However, the difference in the presence or absence of high frequencies in the top layer is less evident for BN than for pink noise. Intermittent sounds make it harder to detect the difference in timbre than continuous sounds.

4.3 Differences with 3D audio production experience

Since six of the participants had experience in producing 3D audio, such as 22.2 multichannel sound, their answers were extracted. Fig.6 shows that if the listeners have experience producing 3D audio, they are notably more sensitive to changes in high frequencies. Independent sample t-test (Welch's t-test) was performed to examine differences in the mean values of LCF (low cut filter) and HCF (high cut filter) for each stimulus due to differences in 3D production experience. As a result, no significant difference was found in HCF, but a significant difference was seen in LCF ($t = 1.77$, $df = 125.5$, p-value = 0.040). When comparing the differences between individual stimuli, differences were found between ML (p-value = 0.048) and VL (p-value = 0.051). Figure 7 shows the average values of MK and VL comparing the differences between 3D production experience.

According to the comments from experienced listeners, frequency bands of the top layer were cut, which tended to make them feel more distant or less sense of envelopment. It can be estimated that there is a band contributing to the spatial impression in the sound from the top layer.

4.4 Effect of middle range

The above results demonstrate that the frequency required for the upper layer is a minimum of 400 Hz. According to previous research, a minimum of 4 kHz was necessary for vertical localization; thus, the midrange from 400 Hz to 4 kHz may be related to a spatial impression other than localization. The comments regarding the difference between original and filtered stimuli suggest that it is related to the sense of spaciousness of the top layer. Figure 8 , 9, 10, 11, 12, 13, 14, 15 show the frequency response of the entire stimuli used in the experiments recorded by a dummy head (HATS, B&K4128C) at the center listening position. The blue line in the figure indicates the original, the red line indicates the LCF of 15.7 kHz, and the green line indicates the HCF of 62.5 Hz.

As shown in the figures, there are no noticeable differences from the original frequency responses, even if the band of the top layer is limited. In this experiment, the participants were allowed to move their heads. It is considered that a human perceives a difference in the arrival direction of several sounds that do not appear in the recording of the dummy head by moving the head.

5 Summary

In this study, the authors investigated the difference between the original 22.2 multichannel sound and its filtered sound by limiting the playback frequency band

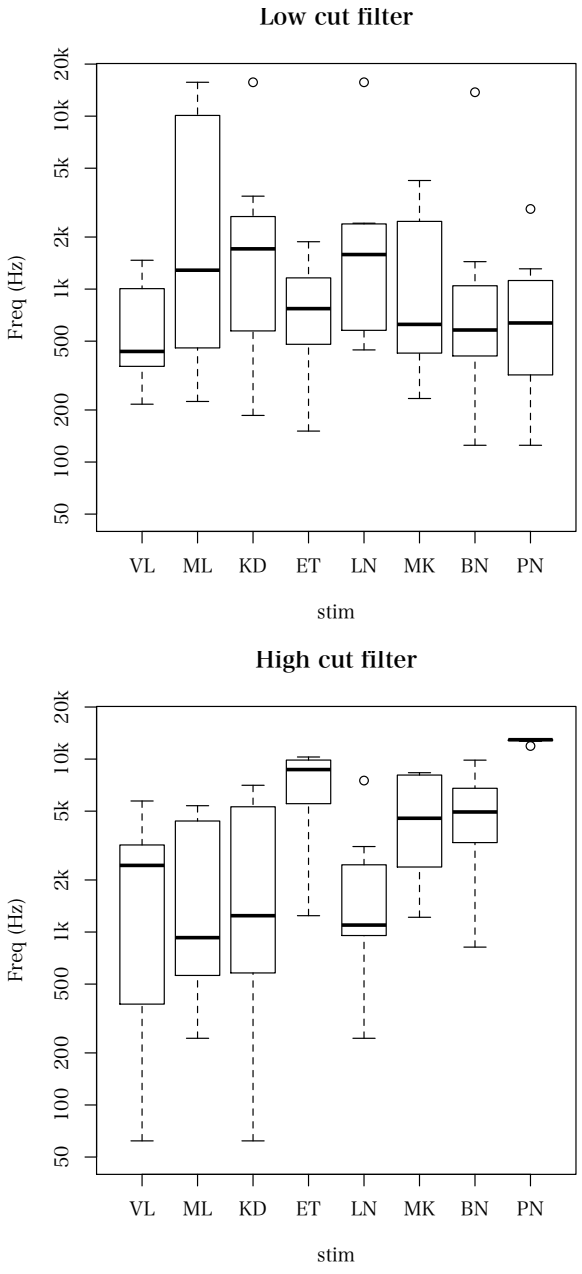

Fig. 6: Boxplots of the low cutoff frequency (upper panel) and high cutoff frequency (lower panel) of the top layer for each stimulus of the six participants who have experience producing 3D audio

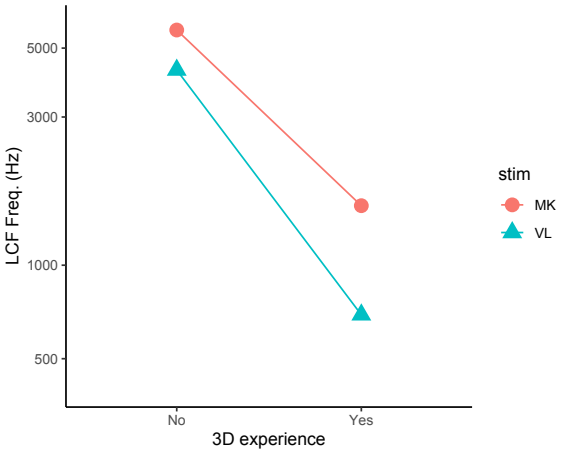

Fig. 7: The average values of MK and VL comparing the differences between experienced and non-experienced 3D production experience.

of its top layer using various contents. In the experiment, the upper and lower ranges of the top layer were cut using the music of various genres. Recordings were made in halls, studios, and with pink noise, etc., and mixed with the middle layer. The difference between the original sound and the impressions were described in words. Subsequently, the participants found the cutoff frequency where the difference between the original sound and the filtered sound could not be distinguished.

As a result, if low frequencies were cut, 75% of listeners did not distinguish the difference from the original sound source if there was a band of 400 Hz or higher. In other words, there is no difference even if the top channel does not have a band below 400 Hz. However, if the high frequency is cut, the sound source and the listener's comments greatly vary. For a sound source including the entire band, such as pink noise, the listener might notice a slight difference in the high frequency.

The above results indicate that the high-frequency band is important for the top layer of the 3D audio, and there is no obstacle to the reproduction of the content even if there is no low-frequency band below 400 Hz. Furthermore, from previous research, the frequency band of 4 kHz or higher is important for sound image localization of vertical direction. However, from

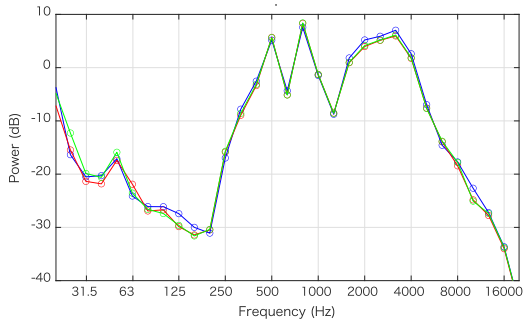

Fig. 8: Frequency response of 1/3 octave band of VL recorded by a dummy head (HATS, B&K4128C) at the center listening position. The blue line indicates the original, the red line indicates the LCF of 15.7 kHz, and the green line indicates the HCF of 62.5 Hz.

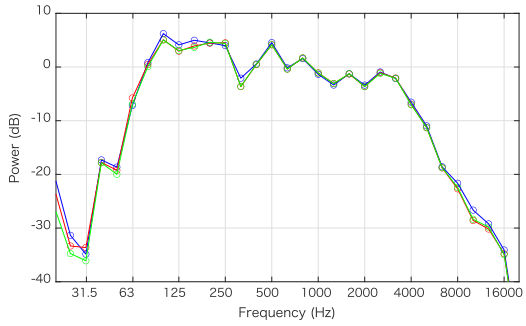

Fig. 9: Frequency response of 1/3 octave band of ML

this result, it was suggested that the frequency band lower than 4 kHz might contribute to spatial impressions such as spaciousness and envelopment of the vertical direction.

From the frequency responses of the stimuli recorded by a dummy head, the sounds with the top layer filtered were not significantly different from the original. The authors estimate that a human perceives a difference in the vertical spatial impression brought from the top layer by moving the head.

However, there is still a variation in the responses among the listeners and contents. If the listeners had experience producing 3D audio, they were notably more sensitive to changes in high and low frequencies. The authors will conduct a more detailed study of the

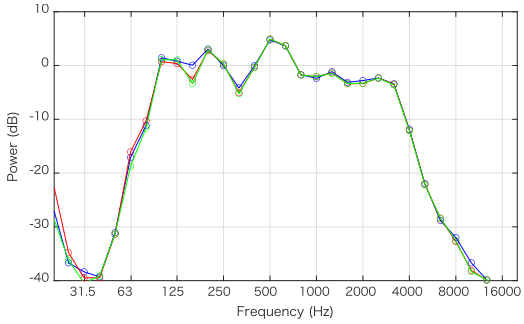

Fig. 10: Frequency response of 1/3 octave band of KD

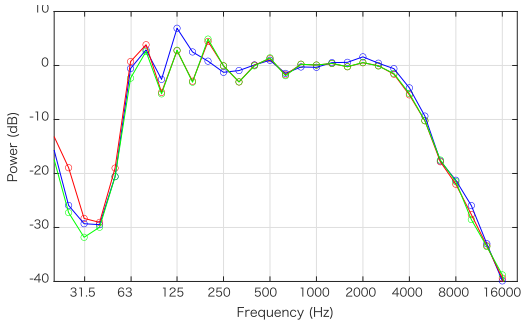

Fig. 11: Frequency response of 1/3 octave band of ET

relationship between frequency band and vertical spatial impression.

References

[1] K. Hamasaki, K. Hiyama, *et al.*, "Advanced Multichannel Audio System with Superior Impression of Presence and Reality," AES 116th Convention, Berlin, Germany,Convention paper 6053, 2004.

[2] Recommendation ITU-R BS.2159-4 "Multichannel sound technology in home and broadcasting applications," International Telecommunication Union, 2012.

[3] Recommendation ITU-R BS.2051-0 "Advanced sound system for programme production," International Telecommunication Union, Geneva, 2014.

[4] L. Rayleigh, "On our perception of sound direction." Philosophical Magazin Series, 6(13), pp. 214-232, 1907.

[5] J. Blauert, "Spatial Hearing," MIT Press, 1997.

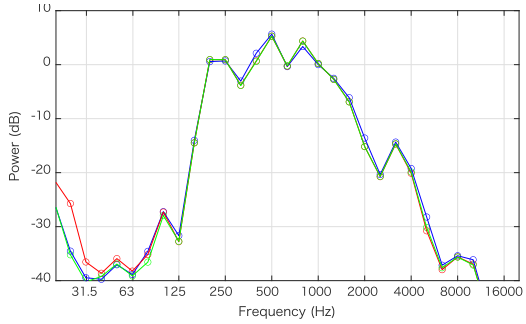

Fig. 12: Frequency response of 1/3 octave band of LN

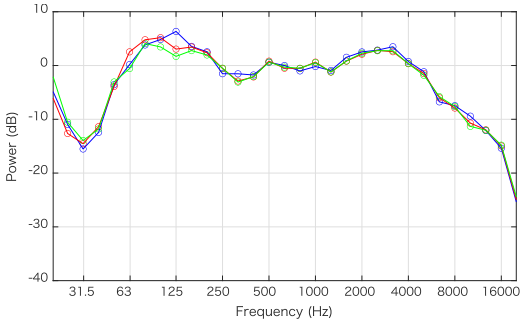

Fig. 14: Frequency response of 1/3 octave band of BN

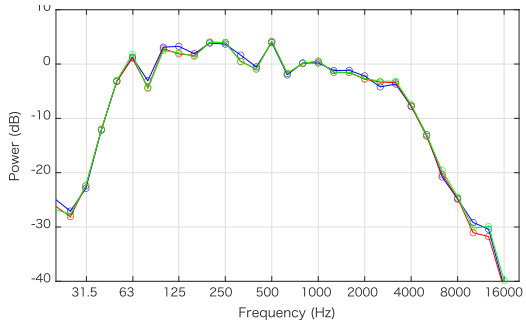

Fig. 13: Frequency response of 1/3 octave band of MK

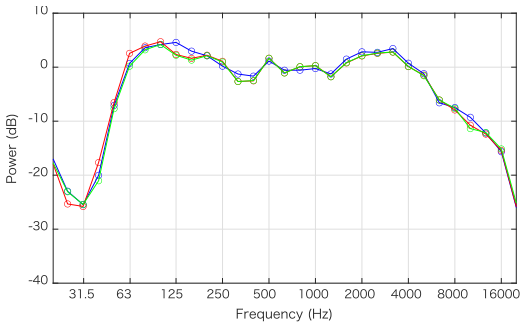

Fig. 15: Frequency response of 1/3 octave band of PN

[6] J. Blauert, "Sound localization in the median plane," ACOUSTICA, 22, pp. 205-213, 1969/1970.

[7] J. Hebrank, D. Wright, "Are two ears necessary for localization of sound sources on the median plane?" Journal of the Audio Engineering Society, 56(3), pp. 935-938, 1974.

[8] M. Morimoto, M. Yairi, K. Iida, M. Ito, "The role of low frequency components in median plane localization," Acoustic Society and Technology, 24(2), 2003.

[9] T. Kamekawa, A. Marui, T. Hosoya. K. Kimura, "Changes in vertical localization and apparent source width of the sound image due to differences in speaker position height," Music Acoustics, Acoustic Society of Japan, (4), 2019 (in Japanese).

[10] M. Morimoto, Z. Maekawa, H. Fujimori, "Discrimination between auditory source width and envelopment," J Acoustic Society Japan, 46(6), pp. 449-457, 1990 (in Japanese).

[11] S. Ferguson, D. Cabrera, "Vertical Localization of Sound from Multiway Loudspeakers," Journal of the Audio Engineering Society, 53(3), 2005.

[12] H. Lee, "Investigationon the PhantomImage Elevation Effect," AES Convention paper, (9441), 2015.

[13] H. Lee, C. Gribben, "Effect of Vertical Microphone Layer Spacing for a 3D Microphone Array," J. Audio Engineering Society, 62(12), pp. 870-884, 2014.

[14] T. Hanyu, K. Hosh and R. Sato, "Spatial impression from late overhead reflections in concert hall," J Acoustic Society of Japan, 69(1), pp. 7-15, 2013.

[15] T. Kamekawa, A. Marui, "Evaluation of recording techniques for three dimensional audio recordings," Acoustic Society and Technology, 41(1), pp. 260-268, 2020.

[16] W. L. Martens, Y. Han, "Discrimination of auditory spatial diffuseness facilitated by head rolling while listening to 'with-height' versus 'without-height',multichannel loudspeaker reproduction,"

AES Conference Paper P4-3, International Conference on Spatial Reproduction - Aesthetics and Science, 2018.