PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

Evaluation of moving sound image localization for reproduction of 22.2 multichannel audio using up-mix algorithm

Hiromu SUZUKI[1]; Takanobu NISHIURA[2]

[1] Graduate School of Information Science and Engineering, Ritsumeikan University, Japan

[2] College of Information Science and Engineering, Ritsumeikan University, Japan

ABSTRACT

22.2 multichannel audio had been developed for ultra-high definition televisions. It can simultaneously reproduce a three-dimensional (3-D) sound field. However, it is difficult to easily acquire multichannel signals because of equipment costs and so on. The up-mix algorithm for generating multichannel signals from the existing fewer channel signals has been proposed to solve this problem. Whereas, the conventional up-mix algorithm has been proposed without the 3-D sound source location. As a result, it cannot reproduce the moving sound image in vertical and backward directions. In this paper, we propose a novel up-mix algorithm which extends two channel signals to 22.2 multichannel signals while using the 3-D location of the sound source. Both objective and subjective evaluations were carried out to compare with the accuracy of reproduced sound image location. Results of the evaluations show that the proposed up-mix algorithm can reproduce 3-D moving sound images more accurately than the conventional one.

Keywords: 22.2 multichannel audio, 3-D sound field, moving sound image

1. Introduction

New sound field reproduction systems are necessary along with the development of video technology. 22.2 multichannel audio [1] is a new reproduction system in conformity with ultra-high definition televisions. In the conventional system such as 5.1 multichannel audio, all loudspeakers are set on the horizontal plane. Therefore, it cannot reproduce a three-dimensional (3-D) moving sound image. 22.2 multichannel audio sets the loudspeakers in three layers: upper, middle and lower. So, it can reproduce 3-D moving sound image.

In 22.2 multichannel audio, the sounds which are recorded by microphones corresponding to each direction are reproduced from each loudspeaker. However, this production method is not suitable because of the high cost of recording. To solve this problem, we use the up-mix algorithm for generating multichannel signals from the existing fewer channel signals. Whereas, the conventional up-mix algorithm has been proposed without the 3-D sound source location. As a result, it cannot reproduce the moving sound image in vertical and backward directions. In this paper, we propose a novel up-mix algorithm which extends two channel signals to 22.2 multichannel signals while using the 3-D location of the sound source. In addition, we try to reproduce 3-D moving sound images with 22.2 multichannel audio using a novel up-mix algorithm.

2. 22.2 multichannel audio

2.1 Design

Fig.1 shows the design of 22.2 multichannel audio. This system can divide three layers: upper, middle and lower. It consists of 9 loudspeakers in the upper layer, 10 loudspeakers in the middle layer, 3 loudspeakers and 2 woofers in the lower layer. These layers can reproduce a 3-D sound field. Fig.2 shows the label of loudspeakers and the installation interval in each layer. This design is based on the conventional study about the optimal installation interval of loudspeakers.

[1] is0312ps@ed.ritsumei.ac.jp
[2] nishiura@is.ritsumei.ac.jp

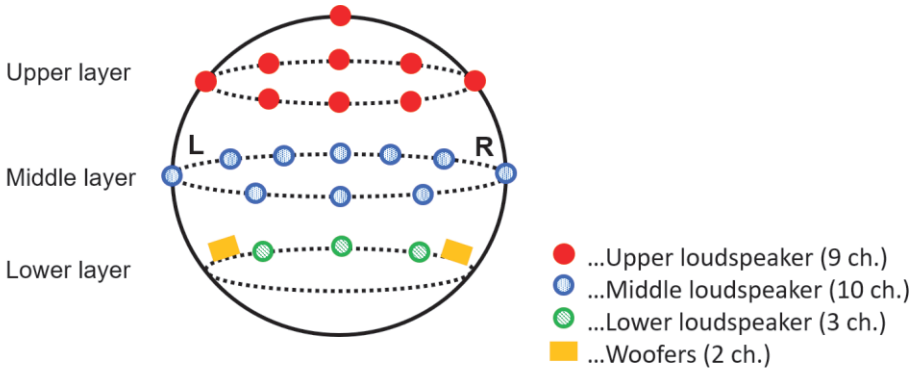
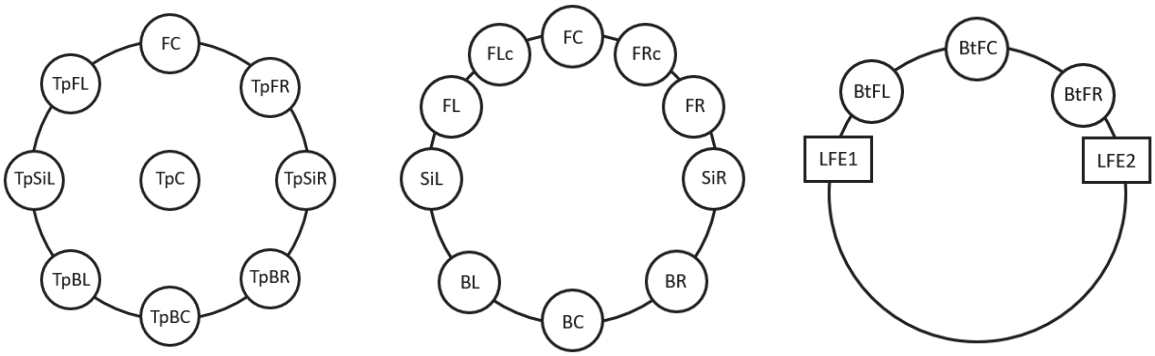


Figure 1 - Design of 22.2 multichannel audio.



(a) Upper layer (b) Middle layer (c) Lower layer

Figure 2 - The label of loudspeakers and the installation interval.

2.2 Effects

The listener obtains the following effects [2] for sound field reproduction with the design shown in Section 2.1.

i. The arrival of sound from all directions
ii. Sound localization in any positions
iii. Construction of high-quality 3-D sound field by reverberation
iv. Matching of video images and sound images

These effects greatly contribute to high sound field reproduction.

2.3 Problem

Driving 22.2 multichannel audio needs as many as 24 channel signals. One of the ways to generate these signals is a recording method using 22 spherical microphones. However, this method is costly because it records sound sources using individually microphones. To solve this problem, we focus on up-mix algorithm: generating multichannel signals from the existing fewer channel signals. In this paper, we propose a novel up-mix algorithm which extends two channel signals to 22.2 multichannel signals while using the 3-D location of the sound source. In addition, we challenge the construction of a 3-D moving sound image using generated signals.

3. Conventional up-mix algorithm based on reverberation separation

3.1 Overview

Multi-channel audio requires faithful reproduction of the direct sound and the reverberation. The reproduced sound should be separated into the direct sound and the reverberation because the reverberation comes from all directions in any places. Therefore, we consider applying conventional up-mix algorithm [3] based on the reverberation separation to 22.2 multichannel audio.

Fig.3 shows the overview of converting two channel signals to 22.2 channel signals using conventional up-mix algorithm. The procedure is following.

i. Reverberation separation

Two channel signals $s_1(t)$ and $s_2(t)$ are separated into direct sound $x_{\mathrm{Direct}}(t)$ and reverberation $x_{\mathrm{Rev}}(t)$ using reverberation separation based on Multi-Step Linear Prediction (MSLP) [4].

ii. Frame division

In this paper, direct sound $x_{\mathrm{Direct}}(t)$ and reverberation $x_{\mathrm{Rev}}(t)$ divided into f frame to reproduce 3-D moving sound image. Divided sounds $x_{\mathrm{Direct}}(t',f)$ and $x_{\mathrm{Rev}}(t',f)$ are reproduced from the loudspeaker for each frame; t' is the time index in f frame.

iii. 2-D Amplitude panning (direct sound)

Direct sound $x_{\mathrm{Direct}}(t',f)$ is mapped as 2-D sound image in the left or right using 2-D amplitude panning. 2-D amplitude panning is achieved by controlling the amplitudes of two channel signals based on added 2-D sound image localization. The coefficients $g_{\mathrm{Conv},1}(f)$ and $g_{\mathrm{Conv},2}(f)$ to control two channel signals are calculated as follows:

$$g_{\mathrm{Conv},1}(f) = \frac{\sin\phi + \sin\theta_f}{\sqrt{2\left(\sin^2\phi + \sin^2\theta_f\right)}}, \tag{1}$$

$$g_{\mathrm{Conv},2}(f) = \frac{\sin\phi - \sin\theta_f}{\sqrt{2\left(\sin^2\phi + \sin^2\theta_f\right)}}, \tag{2}$$

where ϕ is the angle between the listener and the loudspeaker, θ_f is the angle between the listener and the sound image at f frame, $i = 1$ denotes the left channel, $i = 2$ denotes the right channel. Finally, the two channel signals $y_{\mathrm{Direct},1}(t',f)$ and $y_{\mathrm{Direct},2}(t',f)$ are calculated as follows:

$$y_{\mathrm{Direct},i}(t',f) = x_{\mathrm{Direct}}(t',f)\,g_{\mathrm{Conv},i}(f). \qquad \text{s.t. } i = 1,2 \tag{3}$$

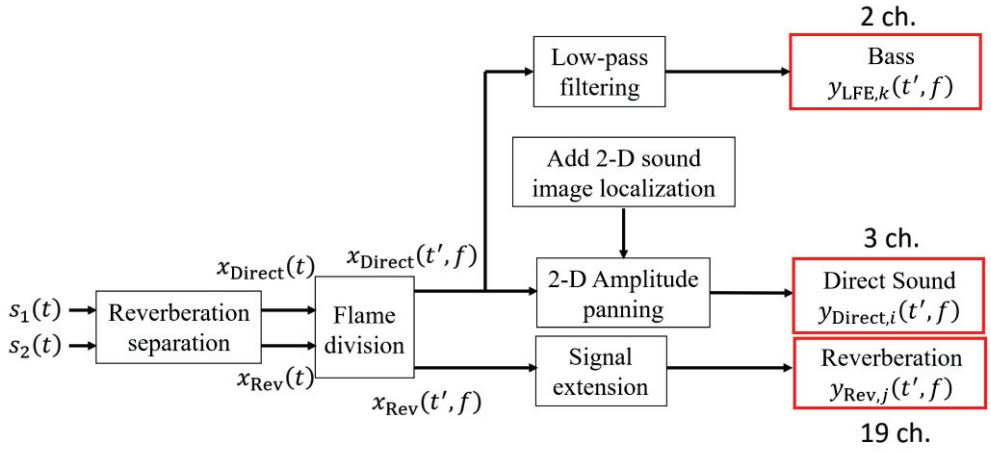


Figure 3 - Conventional up-mix algorithm.

Furthermore, the center channel signal $y_{\mathrm{Direct},3}(t',f)$ applies the direct sound $x_{\mathrm{Direct}}(t',f)$ without panning:

$$y_{\mathrm{Direct},i}(t',f) = x_{\mathrm{Direct}}(t',f). \quad \text{s.t. } i = 3 \tag{4}$$

The three generated signals $y_{\mathrm{Direct},i}(t',f)$ map specific loudspeakers shown in Fig.4.

iv. Signal extension (reverberation)

Reverberation $x_{\mathrm{Rev}}(t',f)$ comes from all directions in any spaces. Therefore, it should be reproduced by remaining 19 loudspeakers. The generated 19 reverberations $y_{\mathrm{Rev},j}(t',f)$ are shown as follows:

$$y_{\mathrm{Rev},j}(t',f) = x_{\mathrm{Rev}}(t',f), \quad \text{s.t. } j = 1,\dots,19 \tag{5}$$

where j shows loudspeaker channel to reproduce the reverberation. The 19 generated signals $y_{\mathrm{Rev},j}(t',f)$ mainly map the rear loudspeakers in Fig.4.

v. Low-pass filtering (bass)

The bass reproduced from the woofer is generated by the low-pass filter. The standard for 22.2 multichannel audio [3] defines to play less than 120 Hz as bass. The two basses $y_{\mathrm{LFE},k}(t',f)$ apply low-pass filter as follows:

$$y_{\mathrm{LFE},k}(t',f) = x_{\mathrm{Direct}}(t',f) * \tau_{\mathrm{LPF}}(t',f), \quad \text{s.t. } k = 1,2 \tag{6}$$

where k shows woofer channel, $\tau_{\mathrm{LPF}}(t',f)$ is the coefficient of the low-pass filter.

3.2 Problem

The conventional up-mix algorithm separates two channel signals into direct sound and reverberation; direct sound is mapped three front loudspeakers, reverberation is mapped remaining 19 loudspeakers. However, the following two problems occur when trying to apply the generated signals to the 22.2 multichannel audio.

i. The difficulty of reproducing the sound field moving up and down

22.2 multichannel audio needs to reproduce 3-D moving sound image in any spaces. Whereas, the conventional up-mix algorithm has been proposed without the 3-D sound source location. We should apply 3-D sound image localization method using upper and lower loudspeakers.

ii. The need to generate a large number of reverberations

In the actual sound field, different reverberations come from all directions due to the reflection of the wall and so on. Whereas, the conventional up-mix algorithm generates the same reverberations to play 19 loudspeakers. We should adopt a method to generate multiple reverberations corresponding to 22.2 multichannel audio.

In this paper, we propose a novel up-mix algorithm using 3-D location of the sound source.

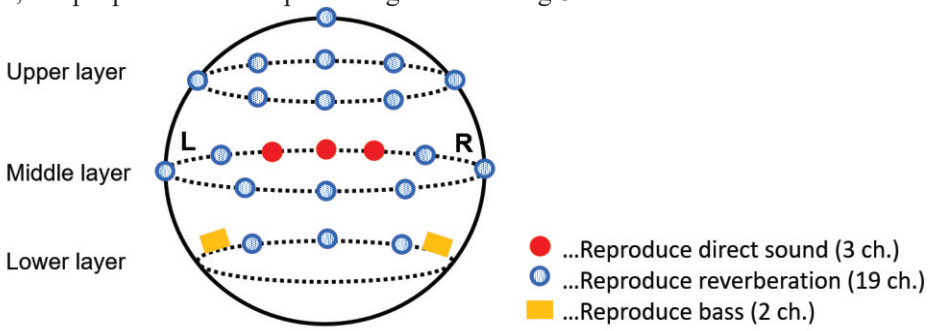


Figure 4 - The mapping method using conventional up-mix algorithm.

4. A novel up-mix algorithm using the 3-D location of the sound source

A novel up-mix algorithm can reproduce a free moving sound image using the 3-D location of the sound source. Fig.5 shows the overview of converting two channel signals to 22.2 channel signals using a novel up-mix algorithm. The procedure is following.

i. Reverberation separation

Two channel signals $s_1(t)$ and $s_2(t)$ are separated into direct sound $x_{\mathrm{Direct}}(t)$ and reverberation $x_{\mathrm{Rev}}(t)$ in the same manner as conventional up-mix algorithm.

ii. Select loudspeaker

This algorithm determines the loudspeakers for reproducing direct sound or reverberation based on added 3-D location of the sound source. The 3-D localization (r, θ_f, ϕ_f) is acquired automatically or manually from media such as video; r is the distance from listening point to loudspeaker, θ_f is the azimuth from listening point to sound source location at f frame, ϕ_f is the elevation from listening point to sound source location at f frame. Direct sound selects three loudspeakers close in distance to the sound source location at f frame. The distance d_i between sound source location and loudspeaker is calculated as follows:

$$d_i = \sqrt{2r^2 - 2r^2 \cos(\phi_i' - \phi_f) \sin\theta_f \sin\theta_i' + \cos\theta_f \cos\theta_i'}, \tag{7}$$

where θ_i' is the azimuth from listening point to loudspeaker at f frame, ϕ_i' is the elevation from listening point to loudspeaker at f frame. The exception is that the all selected loudspeakers should not be in the same layer.

iii. Frame division (direct sound)

The direct sound $x_{\mathrm{Direct}}(t)$ is divided into f frame after loudspeaker selection in Step ii. Divided direct sound $x_{\mathrm{Direct}}(t', f)$ is localized in 3-D sound field by Vector Base Amplitude Panning (VBAP) [5] in Step iv.

iv. VBAP (direct sound)

VBAP is one of the 3-D amplitude panning method by vector synthesis, this method realizes any 3-D sound localizations. Fig.6 shows that VBAP divides the reproduction space into triangle consisting of three loudspeakers and pans amplitude based on the calculated weighting factors $\mathbf{g} = (g_{\mathrm{Prop},1}\ g_{\mathrm{Prop},2}\ g_{\mathrm{Prop},3})^T$. The weighting factors $\mathbf{g}$ are calculated as follows:

$$\mathbf{g} = \mathbf{L}^{-1}\mathbf{p}, \tag{8}$$

where $\mathbf{p}$ is unit vector from the listening point to virtual sound image, $\mathbf{L} = (\boldsymbol{l_1}\ \boldsymbol{l_2}\ \boldsymbol{l_3})$ is matrix of unit vectors $\boldsymbol{l_1}, \boldsymbol{l_2}, \boldsymbol{l_3}$ from the listening point to each loudspeaker. The weighting factors $g_{\mathrm{Prop},1}, g_{\mathrm{Prop},2}, g_{\mathrm{Prop},3}$ for f frame are obtained by Eq. (8) calculated for each frame.

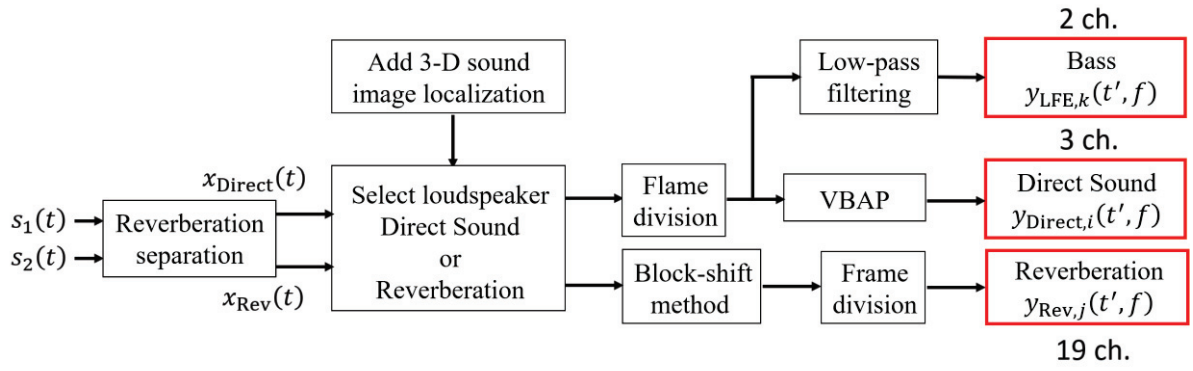


Figure 5 - The mapping method using a novel up-mix algorithm.

The panned sounds $y_{\mathrm{Direct},i}(t',f)$ are the product of the calculated weighting factors $g_{\mathrm{Prop},1}, g_{\mathrm{Prop},2}, g_{\mathrm{Prop},3}$ and the direct sound $x_{\mathrm{Direct}}(t',f)$ as follows:

$$y_{\mathrm{Direct},i}(t',f) = x_{\mathrm{Direct}}(t',f)g_{\mathrm{Prop},i}(f). \qquad \text{s.t. } i = 1,\dots,3 \tag{9}$$

v. Block shift method (reverberation)

Block shift method [6] can artificially generate many impulse responses from an original impulse response. Generated impulse responses are similar to the original one and low correlation between them. We can generate multiple reverberations by convoluting the impulse responses generated by block shift method and the original reverberation. The flow of reverberation generation using the block shift method is shown as follows.

First, original impulse response is divided into multiple n blocks $r_n(t)$ by time windows $w(t)$. $r_n(t)$ is generated as follows:

$$r_n(t) = w\big(t-(n-1)\big)T_d r_A(t), \tag{10}$$

where T_d is the interval between two adjacent time windows, $r_A(t)$ is reverberation part in the original impulse response.

Second, each divided block is randomly shifted and replaced by a new block to generate a new reverberation part $r_B(t)$. $r_B(t)$ is generated as follows:

$$r_B(t) = \sum_{n=1}^{N} r_{n+N^*(n)}(t + N^*(n)T_d), \tag{11}$$

where N is the total number of n blocks, $N^*(n)$ is a random number with an uniform distribution ranging from $-\frac{N}{2}$ to $\frac{N}{2}$.

Finally, we can obtain new 19 channel impulse responses $h_{B,j}(t)$ by combining the direct sound part with the reverberation part $r_B(t)$. New 19 channel reverberations $y_{\mathrm{Rev},j}(t)$ are calculated as follows:

$$y_{\mathrm{Rev},j}(t) = x_{\mathrm{Rev}}(t) * h_{B,j}(t). \qquad \text{s.t. } j = 1,\dots,19 \tag{12}$$

vi. Frame division (reverberation)

The generated reverberations $y_{\mathrm{Rev},j}(t)$ are divided into f frame. Divided reverberations $y_{\mathrm{Rev},j}(t',f)$ map 19 loudspeakers except for three loudspeakers what reproduce direct sound.

vii. Low-pass filtering (bass)

The basses $y_{\mathrm{LFE},k}(t',f)$ for two woofers are generated by the low-pass filter in the same manner as conventional up-mix algorithm.

22.2 multichannel signals generated by the above steps can reproduce a 3-D moving sound image.

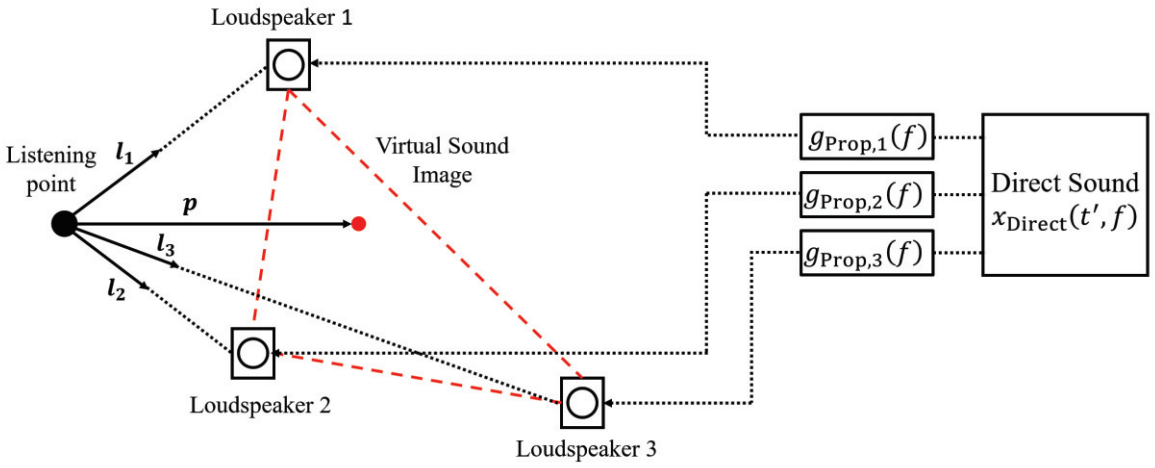


Figure 6 - Vector Base Amplitude Panning.

5. Evaluation experiment

We carried out the objective evaluation experiment to confirm the effectiveness of the proposed method.

5.1 Conditions of objective evaluation experiment

Table 1 shows the experimental conditions. Table 2 shows the experimental equipment. In this experiment, we calculated Interaural Level Difference (ILD) to evaluate the localization direction of the 3-D moving sound image. This objective experiment calculates the ILD of three moving sound sources; measured sound source, evaluation sound source reproduced by conventional up-mix algorithm, evaluation sound source reproduced by proposed up-mix algorithm. Furthermore, we identify localization error by calculating the average of ILD errors E_{ILD} between measured sound source and evaluation sound source. E_{ILD} is calculated as follows:

$$E_{\mathrm{ILD}} = \sqrt{\frac{1}{F}\sum_{f=0}^{F-1}\left(I_{\mathrm{Eva}}(f) - I_{\mathrm{Real}}(f)\right)^2}, \tag{13}$$

$$I_{\mathrm{Eva}}(f) = \left(\frac{\sum_{t'=fM}^{fM+M-1}|R(t',f)|}{\sum_{t'=fM}^{fM+M-1}|L(t',f)|}\right), \tag{14}$$

where F is the total number of f frames, M is frame width, t' is the time index in f frame, $I_{\mathrm{Eva}}(f)$ is the ILD of each evaluation sound source, $I_{\mathrm{Real}}(f)$ is the ILD of each measured sound source, $L(t',f)$ and $R(t',f)$ are received signals recorded by dummy head microphone.

We carried out the experiment using six 3-D movement patterns shown in Table 2.

Table 1 - Experimental conditions in objective evaluation.

Environment	Experiment room
Ambient noise level	L_A = 37.0 dB
Reverberation time	T_{60} = 300 ms
Sampling frequency	48 kHz
Sound source	White noise (3 s)
3-D movement patterns	① from front right to front left ② from back left to back right ③ from front up to front down ④ from back down to back up ⑤ from front left to back left ⑥ from front right to back right

Table 2 - Experimental equipment in objective evaluation.

Dummy Head Microphone	NEUMANN, KU 100
Loudspeaker	YAMAHA, VXS5
Woofer	YAMAHA, VXS10S
Microphone amplifier	THINKNET, MA-2016C
Loudspeaker amplifier	YAMAHA, XMV 8280
A/D, D/A converter	RME, MADIface USB

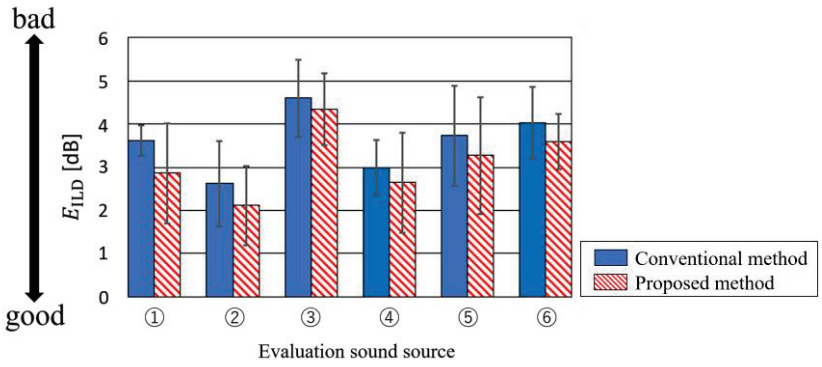


Figure 7 - The average of ILD errors.

5.2 Experimental results in objective evaluation

Fig.7 shows the experimental results. Horizontal axis represents evaluation sound source, and left vertical axis represents the average of ILD errors E_{ILD} between measured sound source and evaluation sound source. From Fig.8, we confirmed that the average of ILD errors decreased in all sound sources. This result suggests that the 22.2 multichannel signals generated by novel up-mix algorithm can reproduce a 3-D moving sound image similar to a real sound source.

6. Conclusions

In this paper, we proposed a novel up-mix algorithm which extends two channel signals to 22.2 multichannel signals while using the 3-D location of the sound source. We confirmed the effectiveness of the proposed method from an evaluation experiment with interaural level distance. As future works, we challenge the reproduction of approach sound based on the control of localization distance using a novel up-mix algorithm.

ACKNOWLEDGEMENTS

This work was partly supported by JST COI and JSPS KAKENHI Grant Numbers JP18K19829 and JP19H04142.

REFERENCES

1. K. Hamasaki, T. Nishiguchi, R. Okumura, Y. Nakayama and A. Ando, "A 22.2 Multichannel Sound System for Ultrahigh-Definition TV (UHDTV)," in SMPTE Motion Imaging Journal, vol. 117, no. 3, pp. 40-49, April 2008.
2. Recommendation ITU-R, "Performance requirements for an advanced multichannel stereophonic sound system for use with or without accompanying picture," BS. 775, 2012.
3. K. Kinoshita, T. Nakatani and M. Miyoshi, "Blind upmix of stereo music signals using multi-step linear prediction based reverberation extraction," IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, USA, 2010, pp. 49-52.
4. K. Kinoshita, M. Delcroix, T. Nakatani and M. Miyoshi, "Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction," IEEE Transactions on Audio, Speech, and Language Processing, vol. 17, no. 4, pp. 534-545, May 2009.
5. V. Pulkki, "Uniform spreading of amplitude panned virtual sources, "Proceedings of the 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics," New Paltz, NY, USA, 1999, pp. 187-190.
6. C. Mori, T. Nishiguchi and K. Ono, "Measurement and generation method of many impulse responses for 22.2 multichannel sound production," 2016 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), Nara, 2016, pp. 1-4.