Performance Evaluation of 3D Sound Field Reproduction System with a Few Loudspeakers and Wave Field Synthesis

M. Naoe, T. Kimura, Y. Yamakata and M. Katsumoto

National Institute of Information and Communications Technology, 4-2-1, Nukui-Kitamachi, Koganei, 184-8795 Tokyo, Japan
t-kimura@nict.go.jp

A conventional 3D sound field reproduction system using wave field synthesis places a lot of loudspeakers around the listener. However, since such a system is very expensive and loudspeakers come into the listener's field of vision, it is very difficult to construct an audio-visual system with it. We developed and evaluated a 3D sound field reproduction system using eight loudspeakers placed at the vertex of cube and wave field synthesis. We compared the sound localization of a loudspeaker array with that of seventeen loudspeakers placed around the listener and found that their localization capabilities were equivalent except the normal direction of cube's planes.

1 Introduction

Recently, several sound field reproduction techniques have been developed for auditory virtual reality systems. By practical application of these techniques, people in different places can conduct and participate in events such as conferences (teleconferencing system) and music concerts (tele-ensemble system) at the same time. Thus, it can be stated that the use of telecommunication systems in society will increase rapidly as these systems are capable of producing more realistic environments and sensations than conventional systems (TV phone and 5.1 ch audio).

Wave field synthesis [1–5] is a sound field reproduction technique that synthesizes wave fronts by using Huygens' principle. The original sound is first recorded using a microphone array in a control area and then reproduced in a listening area by a loudspeaker array. The arrays are placed at the boundaries of their respective areas. The positions of the microphones and loudspeakers are the same with regard to their respective areas. This technique enables multiple listeners to move about in a listening area or to turn their heads and still hear the same sound. This type of sound field reproduction is not possible with conventional sound field reproduction techniques such as the binaural [6] and transaural [7] techniques.

In conventional sound field reproduction systems that use wave field synthesis, loudspeakers are placed in a line [1][3] or surround the listener on a horizontal plane [2][4–5] in order to reproduce the sound field of a 2D space. The sound field reproduction system, in which a lot of loudspeakers are placed around the listener, is also proposed in order to reproduce the sound field of a 3D space [8]. However, since these systems are very expensive and the loudspeakers are visible in the listener's field of vision, it is very difficult to construct an audio-visual system using these systems.

The number of microphones and loudspeakers used by the system can be reduced by considering the auditory capability of the listeners, even if wave fronts are reproduced in the low-frequency range [4]. Thus, by performing a listening test and gauging the auditory capability of the listeners, a practical system can be constructed using only the minimum required number of microphones and loudspeakers.

In this study, we investigate the performance of a 3D sound field reproduction system with eight loudspeakers placed at the vertices of a cube. The use of wave field synthesis is proposed to reproduce the 3D sound field, even when the number of loudspeakers is considerably reduced to prevent the loudspeakers from appearing in the listener's field of vision. The auditory capability of the proposed system is evaluated by the localization test.

The diagram of the proposed 3D sound field reproduction system is shown in Figure 1. First, a sound is recorded using a cubic microphone array, as shown on the left in Figure 1. Second, the recorded sound is reproduced by the cubic loudspeaker array, as shown on the right in Figure 1. As a result, the 3D sound field captured by the microphone array is reproduced by the loudspeaker array. Thus, as shown in Figure 1,the listener, who is in the loudspeaker array, feels that sound is moving above their head when sound is moving above the microphone array.

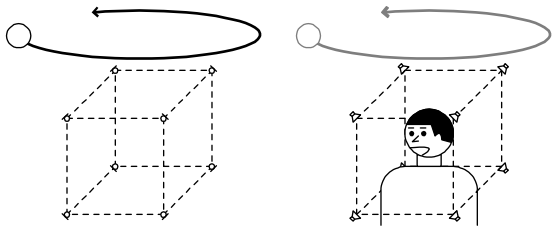

Figure 1 Proposed 3D sound field reproduction system

2 Localization test

In order to evaluate the auditory capability of the proposed 3D sound field reproduction system, a localization test was performed.

2.1 Construction of loudspeaker array

A loudspeaker was manufactured by mounting a loudspeaker unit (Aurasound: modified NSW1-205-8A) on a loudspeaker box, which is designed as shown in Figure 2. The manufactured loudspeaker is shown in Figure 3.

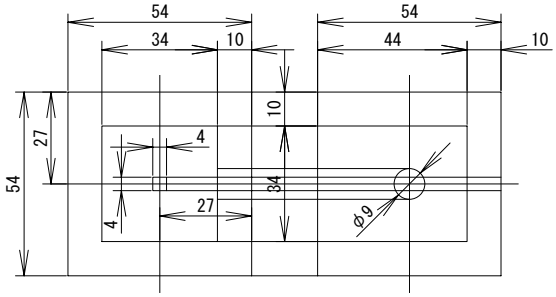

Figure 2 Design of loudspeaker box

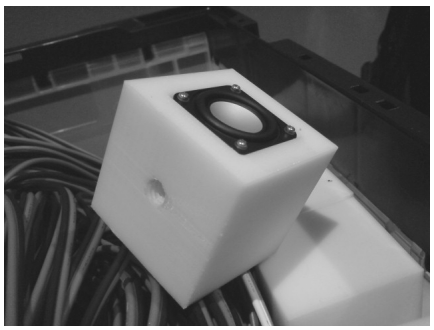

Figure 3 Image of manufactured loudspeaker

Twenty-five loudspeakers were placed in the positions shown in Figure 4. Eight loudspeakers were placed at the vertex of a cube having sides measuring 0.4 m. Seventeen loudspeakers were placed on a sphere with a radius of 1 m; these loudspeakers were used for the control condition. The setup of the loudspeaker array and loudspeakers for the control condition is shown in Figure 5. The white boxes in Figure 5 denote the manufactured loudspeakers.

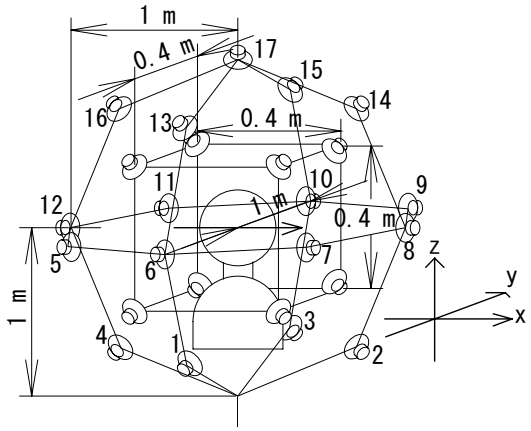

Figure 4 Position of a listener and the loudspeakers

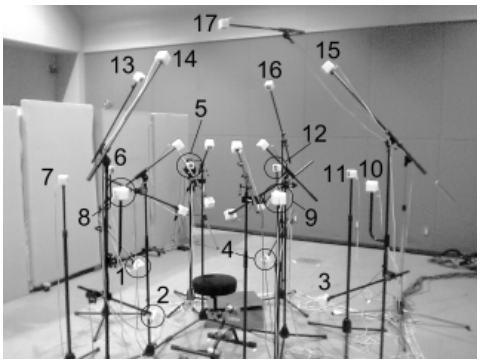

Figure 5 Setup of the loudspeaker array and loudspeakers for the control condition

2.2 Synthesis of multichannel signals

The multichannel signals played by the loudspeaker array were synthesized on a computer. The room impulse response from the sound source to the *i*th microphone (i = 1...8), $g_i(n)$, is denoted as follows:

$$g_i(n) = \frac{1}{d_i}\delta\left\{n - \mathrm{round}\left(\frac{d_i F_s}{c}\right)\right\}, \tag{1}$$

where F_s (=48 kHz) is the sampling frequency, c (=340 m/s) is the sound velocity, $\delta(n)$ is Dirac's delta function, and d_i (=$|\mathbf{r}_0$-$\mathbf{r}_i|$) is the distance between the sound source and the *i*th microphone. The values of $\mathbf{r}_0$ and $\mathbf{r}_i$ (position vectors of the sound source and the *i*th microphone, respectively) were set as follows:

$$\mathbf{r}_0 = \left(d\cos\theta\cos\phi \quad d\sin\theta\cos\phi \quad d\sin\phi\right)^T,$$
$$\mathbf{r}_i = \begin{pmatrix} -0.2 \\ -0.2 \\ -0.2 \end{pmatrix}\begin{pmatrix} 0.2 \\ -0.2 \\ -0.2 \end{pmatrix}\begin{pmatrix} 0.2 \\ 0.2 \\ -0.2 \end{pmatrix}\begin{pmatrix} -0.2 \\ 0.2 \\ -0.2 \end{pmatrix}\begin{pmatrix} -0.2 \\ -0.2 \\ 0.2 \end{pmatrix}\begin{pmatrix} 0.2 \\ -0.2 \\ 0.2 \end{pmatrix}\begin{pmatrix} 0.2 \\ 0.2 \\ 0.2 \end{pmatrix}\begin{pmatrix} -0.2 \\ 0.2 \\ 0.2 \end{pmatrix}, \tag{2}$$

where d (=1, 3 m) denotes the distance between the sound source and the listening position and θ and ϕ are the azimuth and elevation angles, respectively, in the listening position. The values of θ and ϕ were set as shown in Table 1.

If the source signal is represented by $s(n)$, $x_i(n)$, which represents the channel signals recorded by the *i*th microphone, is denoted as follows:

$$\begin{aligned} x_i(n) &= D_i\{g_i(n) * s(n)\} \\ &= \frac{D_i}{d_i} s\left\{n - \mathrm{round}\left(\frac{d_i F_s}{c}\right)\right\}, \end{aligned} \tag{3}$$

where * is the convolution. Previous studies have indicated that the sound is only recorded from outside the control area according to D_i (the directivity of the *i*th microphone) [5]. In this study, D_i was set to shotgun directivity as follows:

Table 1 Azimuth and elevation angles of sound sources

Number	θ [°]	ϕ [°]	Number	θ [°]	ϕ [°]
1	−90	−45	10	90	0
2	0	−45	11	135	0
3	90	−45	12	180	0
4	180	−45	13	−90	45
5	−135	0	14	0	45
6	−90	0	15	90	45
7	−45	0	16	180	45
8	0	0	17	---	90
9	45	0			

$$D_i = \begin{cases} \cos\theta_i & (|\theta_i| \le 90^\circ) \\ 0 & (|\theta_i| > 90^\circ)' \end{cases} \tag{4}$$

where θ_i (incident angle of the sound source in the *i*th microphone) is defined as follows:

$$\theta_i = \cos^{-1}\left\{\frac{\mathbf{r}_i \cdot (\mathbf{r}_0 - \mathbf{r}_i)}{|\mathbf{r}_i||\mathbf{r}_0 - \mathbf{r}_i|}\right\}. \tag{4}$$

On the other hand, signals were also synthesized for the control condition in which only one of the seventeen loudspeakers placed on the sphere of radius 1 m was used for playing the sound.

2.3 Environment and procedure

The localization test was performed in a room with a reverberation time of 180 ms and a background noise level of 23 dB(A). The sound pressure level in the listening position was set to 60 dB(A).

Six males and one female participated as listeners in this study. The flowchart of the localization test is shown in Figure 6. In the test, two sound sources (white noise and speech) were used. Fifty-one stimuli were synthesized for each sound source. The 51 stimuli were divided into sound images of 17 directions and three conditions—the control condition, the condition where the distance between the sound image and the listening position was 1 m (termed 'distance 1 m'), and the condition where the distance was 3 m (called 'distance 3 m'). One hundred fifty-three main trials were performed following thirty-four practice trials were conducted. During the main trials, rest periods were allowed after every set of 51 trials. The orders of the sound sources and the trials were randomised for each listener. The details of the practice and main trials are shown in Table 2.

The listeners were instructed to report the perceived direction of sound by listing the number of the direction in an answer sheet. The relation between the perceived directions and direction numbers is shown in Figure 7. The listeners were allowed to turn their heads freely while listening to the sounds.

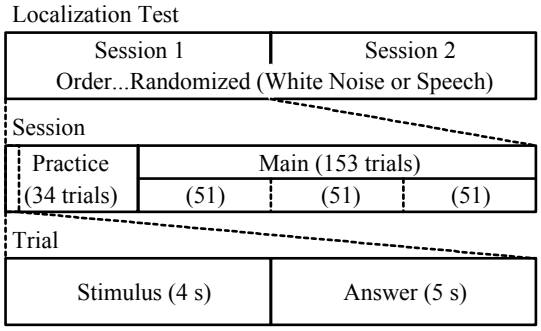

Figure 6 Flow chart of the localization test

Table 2 Details of the practice and main trials

	Element	Note
Practice (34)	= 17 directions × 2 conditions	control, ‘1 m’
Main (153)	= 17 directions × 3 conditions × 3 repetitions	control, ‘1 m’, ‘3 m’

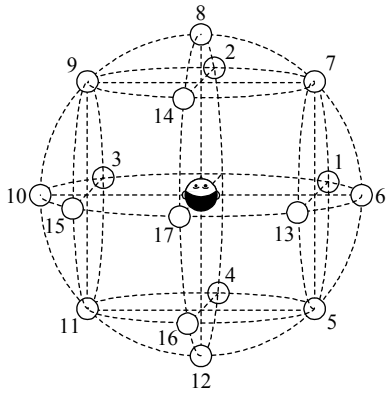

Figure 7 Relation between perceived directions and direction numbers

2.4 Results and discussions

The average rate of accuracy for each condition is shown in Table 3. The accuracy rate was almost 100% for the control condition. This was due to the fact that as the listeners were able to freely move their heads, they could easily localize the sound source. On the other hand, the accuracy rate of the proposed system was approximately 75%. Thus, it can be considered that the localized capability of the proposed system is sufficient, if the capability of this system is judged only by the value of the accuracy rate.

The results of the x^2 test are shown in Tables 4 and 5 and the corresponding accuracy rates for each direction are shown in Figures 8–11. Note that * and ** in Tables denote the significant difference of 5 % and 1 % levels. It was observed that in five directions (6, 8, 10, 12 and 17), the accuracy rates of the proposed system were lower than those of the control condition because there were the differences of the 1% significance level. However, in other directions, the accuracy rates of the proposed system were almost the same as those of the control condition since there were little differences of a 1% significance level. Thus, it is considered that the performance of the proposed system is good in all the directions, except in the five directions stated above.

The results of the answer sheets for the five directions (6, 8, 10, 12, and 17) are shown in Figures 12 and 13 and Tables 6 and 7.

It was observed that in some cases when the direction number was actually 6, an erroneous answer of 7 or 13 was provided. Thus, it is considered that when a sound image is produced from the direction on the right hand side of the listeners, they localize the sound image towards the forward and upper directions.

On the other hand, when the direction number was 8, the most common erroneous answer was 14. Thus, it can be inferred that when a sound image is produced from the region in front of the listeners, the listeners localize the sound image towards the upper direction.

For the direction number 10, the erroneous answer was 3. This implies that the listeners localize a sound image towards the downward direction when the sound image is produced from the direction on the left hand side of the listener.

Table 3 Accuracy rate for each condition

	White Noise	Speech	Average
Control Condition	98%	96%	97%
1m distance	76%	77%	76%
3m distance	76%	74%	75%

Table 4 Results of x^2 test for a white noise

Number	Control condition	1m distance	3m distance
1	100%	81%*	76%*
2	95%	100%	95%
3	100%	81%*	86%
4	95%	76%	90%
5	100%	100%	100%
6	100%	43%**	52%**
7	100%	100%	100%
8	100%	33%**	43%**
9	100%	100%	100%
10	100%	52%**	43%**
11	100%	100%	95%
12	100%	48%**	33%**
13	100%	90%	90%
14	95%	100%	100%
15	100%	76%*	86%
16	86%	71%	71%
17	100%	38%**	33%**

Table 5 Results of x^2 test for speech

Number	Control condition	1m distance	3m distance
1	100%	86%	81%*
2	86%	76%	86%
3	100%	100%	100%
4	90%	81%	71%
5	95%	95%	95%
6	100%	43%**	57%**
7	100%	100%	90%
8	100%	67%**	62%**
9	100%	76%*	90%
10	100%	57%**	52%**
11	100%	90%	90%
12	95%	76%	33%**
13	100%	76%*	81%*
14	100%	95%	95%
15	95%	71%*	71%*
16	86%	67%*	52%**
17	90%	48%**	52%*

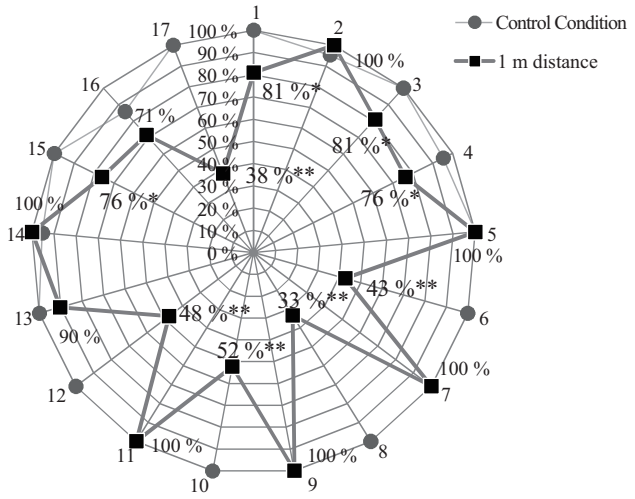

Figure 8 Accuracy rates for each direction when a white noise is used (‘1m distance’ condition)

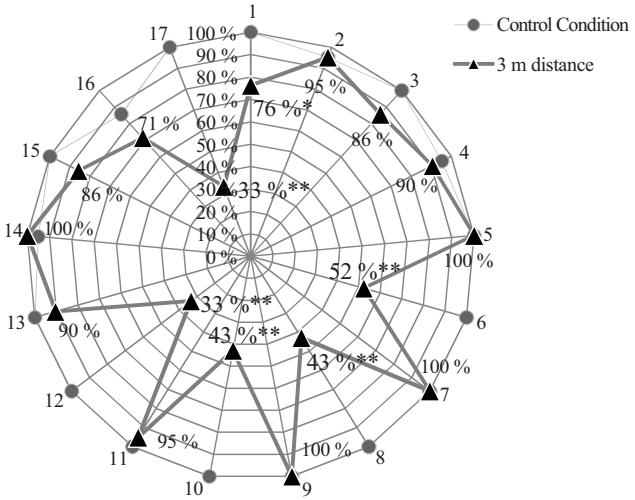

Figure 9 Accuracy rates for each direction when a white noise is used (‘3m distance’ condition)

In the case when the direction number was 12, the erroneous answers were 4, 16, and 17. Thus, it can be inferred that when a sound is produced behind the listeners, they localize the sound image towards the upper direction. Moreover, in such cases, blurring of the sound image can also occur.

When the direction number was 17, the most common erroneous answer was 14. Thus, it can be considered that the listeners localize a sound image towards the forward direction when the sound image is produced from above the listener.

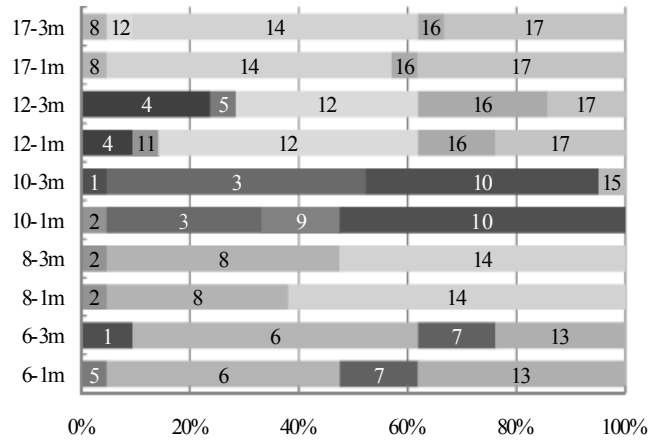

Figure 12 Answers indicating perceived direction of sound for the five directions when white noise is used as a source

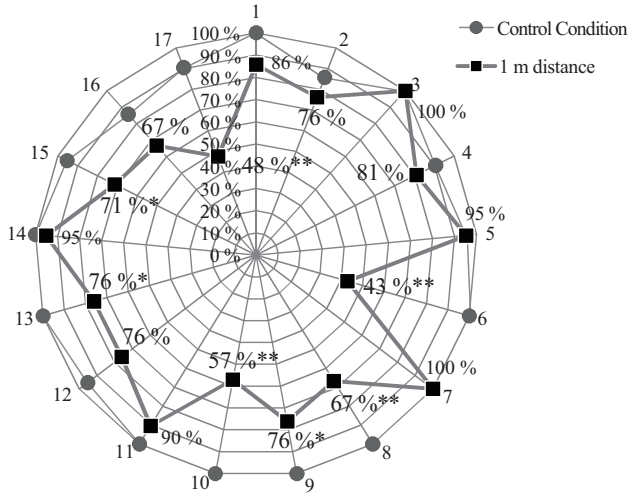

Figure 10 Accuracy rates for each direction when speech is used as a source (‘1m distance’ condition)

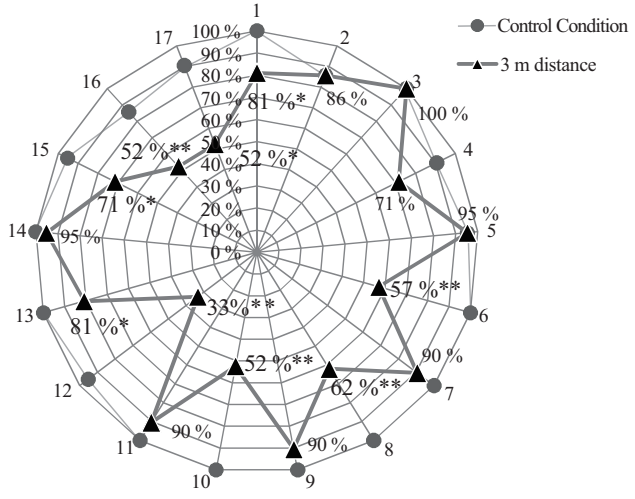

Figure 11 Accuracy rates for each direction when speech is used as a source (‘3m distance’ condition)

It should be noted that identical signals were played from four loudspeakers in all the five directions. Thus, it is considered that the blur and bias in sound images occurred due to phantom sources in the five directions.

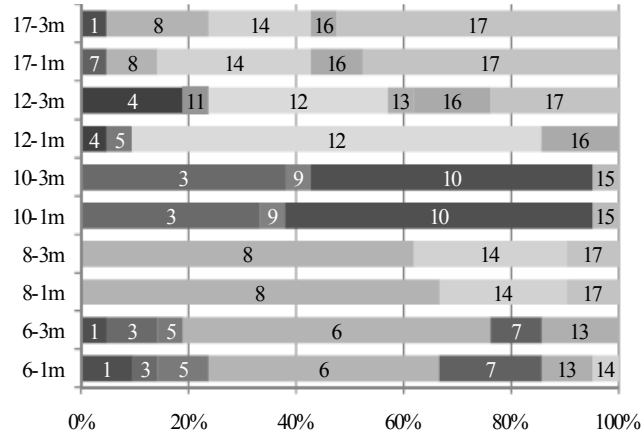

Figure 13 Answer indicating perceived direction of sound for the five directions when speech is used as a source

Table 6 Answer indicating perceived direction of sound for the five directions when white noise is used as a source

Number		6		8		10		12		17	
		1m	3m	1m	3m	1m	3m	1m	3m	1m	3m
Perceived Number	1	0	10	0	0	0	5	0	0	0	0
	2	0	0	5	5	5	0	0	0	0	0
	3	0	0	0	0	29	48	0	0	0	0
	4	0	0	0	0	0	0	10	24	0	0
	5	5	0	0	0	0	0	0	5	0	0
	6	43	52	0	0	0	0	0	0	0	0
	7	14	14	0	0	0	0	0	0	0	0
	8	0	0	33	43	0	0	0	0	5	5
	9	0	0	0	0	14	0	0	0	0	0
	10	0	0	0	0	52	43	0	0	0	0
	11	0	0	0	0	0	0	5	0	0	0
	12	0	0	0	0	0	0	48	33	0	5
	13	38	24	0	0	0	0	0	0	0	0
	14	0	0	62	52	0	0	0	0	52	52
	15	0	0	0	0	0	5	0	0	0	0
	16	0	0	0	0	0	0	14	24	5	5
	17	0	0	0	0	0	0	24	14	38	33

Unit: [%]

Table 7 Answers indicating perceived direction of sound for the five directions when speech is used as a source

Number		6		8		10		12		17	
		1m	3m	1m	3m	1m	3m	1m	3m	1m	3m
Perceived Number	1	10	5	0	0	0	0	0	0	0	5
	2	0	0	0	0	0	0	0	0	0	0
	3	5	10	0	0	33	38	0	0	0	0
	4	0	0	0	0	0	0	5	19	0	0
	5	10	5	0	0	0	0	5	0	0	0
	6	43	57	0	0	0	0	0	0	0	0
	7	19	10	0	0	0	0	0	0	5	0
	8	0	0	67	62	0	0	0	0	10	19
	9	0	0	0	0	5	5	0	0	0	0
	10	0	0	0	0	57	52	0	0	0	0
	11	0	0	0	0	0	0	0	5	0	0
	12	0	0	0	0	0	0	76	33	0	0
	13	10	14	0	0	0	0	0	5	0	0
	14	5	0	24	29	0	0	0	0	29	19
	15	0	0	0	0	5	5	0	0	0	0
	16	0	0	0	0	0	0	14	14	10	5
	17	0	0	10	10	0	0	0	24	48	52

Unit: [%]

3 Conclusion

In this study, we have proposed the 3D sound field reproduction system with eight loudspeakers to reproduce sound. Wave field synthesis allows us to reproduce a 3D sound field even when the number of loudspeakers is made very small in order to prevent the loudspeakers from appearing in the listener's field of vision. The auditory capability of the proposed system was evaluated by the localization test. It was found that the average rate of accuracy of localization was 75% and that a good performance was observed for twelve of the seventeen directions that were used in the test. Moreover, in future studies, we plan to develop a method to improve the localized accuracy of the remaining five directions and test its performance using the localization test.

References

[1] H. Fletcher, "Symposium on wire transmission of symphonic music and its reproduction on auditory perspective: Basic requirement," *Bell Sys. Tech. J.*, vol. 13, no. 2, pp. 239–244, April 1934.

[2] M. Camras, "Approach to recreating a sound field," *J. Acoust. Soc. Am.*, vol. 43, no. 6, pp. 1425–1431, June 1968.

[3] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *J. Acoust. Soc. Am.*, vol. 93, no. 5, pp. 2764–2778, May 1993.

[4] T. Kimura, K. Kakehi, K. Takeda, and F. Itakura, "Subjective assessments for the effect of the number of channel signals on the sound field reproduction used in wavefield synthesis," in *Proc. Int. Cong. Acoust.*, Kyoto, Japan, April 2004, number Th.P1.18, IV, pp. 3159–3162.

[5] T. Kimura and K. Kakehi, "Effects of directivity of microphones and loudspeakers in sound field reproduction based on wave field synthesis," in *Proc. Int. Cong. Acoust.*, Madrid, Spain, September 2007, number RBA-15-011, pp. 1–6.

[6] J. Blauert, *Spatial Hearing*, pp. 372–392, MIT Press, Cambridge, Mass, revised edition, 1997.

[7] J. Bauck and D. H. Cooper, "Generalized transaural stereo and applications," *J. Audio Eng. Soc.*, vol. 44, no. 9, pp. 683–705, September 1996.

[8] S. Ise, M. Toyoda, S. Enomoto, and S. Nakamura, "The development of the sound field sharing system based on the boundary surface control principle," in *Proc. Int. Cong. Acoust.*, Madrid, Spain, September 2007, number ELE-04-003, pp. 1–7.