PAPER

Localization Model of Synthesized Sound Image Using Precedence Effect in Sound Field Reproduction Based on Wave Field Synthesis

Toshiyuki KIMURA[†a)], **Yoko YAMAKATA**[†], **Michiaki KATSUMOTO**[†], ***and*** **Kazuhiko KAKEHI**[††], ***Members***

SUMMARY Although it is very important to conduct listening tests when constructing a practical sound field reproduction system based on wave field synthesis, listening tests are very expensive. A localization model of synthesized sound images that predicts the results of listening tests is proposed. This model reduces the costs of constructing a reproduction system because it makes it possible to omit the listening tests. The proposed model uses the precedence effect and predicts the direction of synthesized sound images based on the inter-aural time difference. A comparison of the results predicted by the proposed model and the localized results of listening tests shows that the model accurately predicts the localized results.

key words: *Sound field reproduction, Wave field synthesis, Localization model, Precedence effect, Inter-aural time difference*

1. Introduction

Sound field reproduction techniques were recently developed for acoustic scene reproduction. If these techniques are practically applied, people in different places can experience conferencing as though they are in the same conference room (teleconferencing system) and play music as though they are in the same concert hall (tele-ensemble system). Since these systems produce more realistic sensations than conventional systems (TV phone and 5.1 ch audio), telecommunication will be more useful and commonly used in society as a whole if these systems are applied.

Wave field synthesis [1]–[3] is a sound field reproduction technique that synthesizes wave fronts based on Huygens' principle. This technique picks up original sound using a microphone array in a control area and then reproduces it in a listening area using a loudspeaker array. The arrays are placed at the boundaries of their respective areas. The positions of the microphones and the 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. Conventional sound field reproduction techniques, such as binaural [4] and transaural [5], cannot do this.

Until recently it was impossible to construct a practical wave field synthesis system. This is because, according to spatial sampling theorem, microphones and loudspeakers should be placed at intervals of less than half the wavelength to reproduce physical wave fronts. However, since listening tests showed that the number of microphones and loudspeakers reproducing realistic sensations can be reduced [6], [7], it has been possible to construct practical systems if listening tests are done.

However, listening tests are expensive because they must be done for each microphone and loudspeaker array shape required by the application. If the results of listening tests can be predicted based on the position of microphones and loudspeakers, fewer tests will need to be done. The number of microphones and loudspeakers reproducing directional perception was more than that of microphones and loudspeakers reproducing spatial impression when the listening tests are done for each realistic sensation parameter [6], [7] (directional perception, distant perception and spatial impression [8]). Thus, it is important to construct a localization model that predicts localized results of listening tests.

In conventional localization models [9]–[12], a direction is predicted based on the inter-aural time difference (ITD) calculated from the input binaural signals. The ITD is also used as the estimation criterion of directions in this paper. The precedence effect [13] must also be introduced to the localization model because perception of direction is biased by the precedence effect when there are only a few microphones and loudspeakers [6], [7].

However, Lindemann's model [10], [11], which introduces the precedence effect to the localization model, does not use the acoustic transfer function between loudspeakers and listener's ears because this model assumes that the listener is listening to a sound by headphones. Kurozumi et al.'s model [12], in contrast, takes account of the acoustic transfer function between two loudspeakers and the listener's ears in the localization model but does not take account of the precedence effect. Thus, no localization model that uses both the acoustic transfer function and the precedence effect has yet been proposed.

We propose a localization model that uses the acoustic transfer function between loudspeakers and listener's ears and the precedence effect to predict localized results of listening tests of sound field reproduction based on wave field synthesis. The algorithm for the proposed model is described in Section 2. How model parameters were set to construct the model is described in Section 3. In Section 4, the predicted results of the proposed model are compared with the localized results of listening tests [6], [7] and the

Manuscript received August 7, 2007.
Manuscript revised November 23, 2007.
Final manuscript received December 21, 2007.

[†]The author is with the Universal Media Research Center, National Institute of Information and Communications Technology, Koganei, Tokyo, 184-8795 Japan.
[††]The author is with the School of Information Science and Technology, Chukyo University, Toyota, Aichi, 470-0393 Japan.
a) E-mail: t-kimura@nict.go.jp

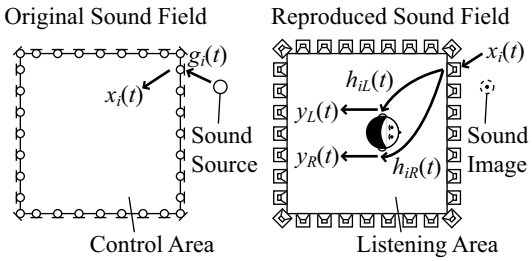


Fig. 1 Synthesis of channel signals and binaural signals in the localization model.

effectiveness of the proposed model is discussed.

2. Algorithm for Localization Model

An original sound field is a free field where there is no reflection sound, since only a direct sound from sound sources that mainly contributes to directional perception. M microphones are placed at the boundary of a control area, as shown in the left side of Fig. 1. The room impulse response from the sound source to the ith microphone, $g_i(t)$, is denoted as follows,

$$g_i(t) = a_i\delta(t - t_i) \quad (i = 1...M), \tag{1}$$

where $a_i(= 1/d_i)$ and $t_i(= d_i/c)$ are the amplitude and the delay depending on distance d_i between the sound source and the ith microphone, c is sound velocity, $\delta(t)$ is Dirac's delta function, and M is the total number of microphones. When the source signal is denoted as $s(t)$, $x_i(t)$ (channel signals recorded by the ith microphone) is denoted as follows,

$$x_i(t) = D_{im}\{g_i(t) * s(t)\} = D_{im}a_i s(t - t_i), \tag{2}$$

where $*$ is the convolution. The sound from the outside of the control area is only recorded based on D_{im} (the directivity of the ith microphone) [14].

In the reproduced sound field, M loudspeakers are placed at the boundary of a listening area, as shown in the right side of Fig. 1. Loudspeakers and microphones are configured in the same way. As in Kurozumi et al.'s model [12], the head-related impulse response (HRIR) from the ith loudspeaker to the listener's left and right ears, $h_{iL}(t)$ and $h_{iR}(t)$, are approximated as follows,

$$\begin{aligned} h_{iL}(t) &\approx a_{iL}\delta(t - t_{iL}), \\ h_{iR}(t) &\approx a_{iR}\delta(t - t_{iR}), \end{aligned} \tag{3}$$

where $a_{iL(iR)}$ and $t_{iR(iL)}$ are the amplitude and the initial delay between the ith loudspeaker and the listener's left and right ears. According to channel signals and HRIRs, the binaural signals of the left and right ears, $y_L(t)$ and $y_R(t)$, are denoted as follows,

$$\begin{aligned} y_L(t) &= \sum_{i=1}^{M} D_{is}\{h_{iL}(t) * x_i(t)\} \\ &= \sum_{i=1}^{M} D_{im}D_{is}a_i a_{iL} s(t - T_{iL}), \\ y_R(t) &= \sum_{i=1}^{M} D_{is}\{h_{iR}(t) * x_i(t)\} \\ &= \sum_{i=1}^{M} D_{im}D_{is}a_i a_{iR} s(t - T_{iR}), \end{aligned} \tag{4}$$

where $T_{iL} = t_i + t_{iL}$, $T_{iR} = t_i + t_{iR}$, and D_{is} is the directivity of the ith loudspeaker. The sound radiates toward the inside of the listening area based on D_{is} [14].

The ITD is calculated from binaural signals. However, the precedence effect isn't introduced in $y_L(t)$ and $y_R(t)$. Thus, it needs to modify the binaural signals in order to introduce the precedence effect as in Lindemann's model [10], [11]. The modified binaural signals, $y'_L(t)$ and $y'_R(t)$, are denoted as follows,

$$\begin{aligned} y'_L(t) &= \sum_{i=1}^{M} p_i D_{is}\{h_{iL}(t) * x_i(t)\} \\ &= \sum_{i=1}^{M} p_i D_{im}D_{is}a_i a_{iL} s(t - T_{iL}), \\ y'_R(t) &= \sum_{i=1}^{M} p_i D_{is}\{h_{iR}(t) * x_i(t)\} \\ &= \sum_{i=1}^{M} p_i D_{im}D_{is}a_i a_{iR} s(t - T_{iR}), \end{aligned} \tag{5}$$

where p_i is the precedence effect coefficient of the ith loudspeaker. p_i is defined as follows,

$$\begin{aligned} p_i &= \exp\{\alpha(t_{\min} - t_i - t'_i)\}, \\ t_{\min} &= \min_i(t_i + t'_i), \end{aligned} \tag{6}$$

where $t'_i(= d'_i/c)$ is the delay depending on distance d'_i between the ith loudspeaker and the listening position and $t_{\min}$ is the arrival time of the shortest path from the sound source to the listening position. Since other arrival times are always longer than $t_{\min}$, p_i is weighted based on those delays so that the ith loudspeaker doesn't contribute to the directional perception. $\alpha(> 0)$ is appropriately defined for the weighting.

The inter-aural correlation function calculated from binaural signals, $R(\tau)$, is denoted as follows,

$$\begin{aligned} R(\tau) &= E\{y'_L(t)y'_R(t - \tau)\} \\ &= \sum_{i=1}^{M}\sum_{j=1}^{M} p_i p_j D_{im}D_{jm}D_{is}D_{js}a_i a_j a_{iL}a_{jR} \\ &\qquad E\{s(t - T_{iL})s(t - T_{jR} - \tau)\}. \end{aligned} \tag{7}$$

Therefore, the contour of the inter-aural correlation function depends on the statistical property of the source signal.

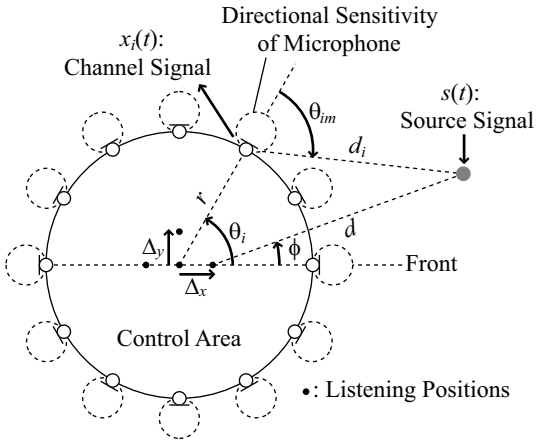


Fig. 2 Synthesis of channel signals in construction of localization model [7].

The only necessary information in the inter-aural correlation function is the peak time because, for our purposes, the ITD is the important information for the model. If the source signal has no auto-correlation property, such as a white noise, to emphasize the peak, Eq. (7) is calculated as follows,

$$R(\tau) = \sum_{i=1}^{M}\sum_{j=1}^{M} P_{ij}D_{ij}A_{ij}\delta(\tau - T_{ij}), \tag{8}$$

where $P_{ij} = p_i p_j$, $D_{ij} = D_{im}D_{jm}D_{is}D_{js}$, $A_{ij} = a_i a_j a_{iL} a_{jR}$, and $T_{ij} = T_{iL} - T_{jR} = (t_i - t_j) + (t_{iL} - t_{jR})$. Since M^2 peaks, which have the amplitude of $P_{ij}D_{ij}A_{ij}$, arise in the position of T_{ij}, as in Kurozumi et al.'s model [12], the effective ITD τ_E is calculated as follows,

$$\tau_E = \frac{\sum_{i=1}^{M}\sum_{j=1}^{M} P_{ij}D_{ij}A_{ij}T_{ij}}{\sum_{i=1}^{M}\sum_{j=1}^{M} P_{ij}D_{ij}A_{ij}}. \tag{9}$$

If the relation between perceived direction ϕ and ITD τ is denoted as follows,

$$\tau = f(\phi), \tag{10}$$

the predicted direction of the synthesized sound image ϕ is calculated from τ_E as follows,

$$\begin{aligned}\phi &= f^{-1}(\tau_E)\\ &= f^{-1}\left(\frac{\sum_{i=1}^{M}\sum_{j=1}^{M} P_{ij}D_{ij}A_{ij}T_{ij}}{\sum_{i=1}^{M}\sum_{j=1}^{M} P_{ij}D_{ij}A_{ij}}\right).\end{aligned} \tag{11}$$

3. Settings of Model Parameters

3.1 Synthesis of Channel Signals

To compare the localized results of listening tests [6], [7], four listening positions (center, front, behind, and lateral) were placed in a circular control area with a radius of two meters, as shown in Fig. 2. Let Δ_x and Δ_y be the moving distance toward the front and left lateral direction from the center of the circle. Then, the coordinates of four listening positions are denoted as follows,

$$(\Delta_x, \Delta_y) = \begin{cases}(0, 0) & \text{(Center)}\\ (0.5, 0) & \text{(Front)}\\ (-0.5, 0) & \text{(Behind)}\\ (0, 0.5) & \text{(Lateral)}\end{cases}, \tag{12}$$

where the units are meters. Let d be the distance between the sound source and the listening position. Then, distance d_i between the sound source and the ith microphone is denoted as follows,

$$\begin{aligned} d_i &= \sqrt{d_x^2 + d_y^2},\\ d_x &= d\cos\phi + \Delta_x - r\cos\theta_i,\\ d_y &= d\sin\phi + \Delta_y - r\sin\theta_i,\end{aligned} \tag{13}$$

where ϕ is the azimuth angle of the sound source in the listening position and θ_i is the azimuth angle of the ith microphone, as shown in Fig. 2. From d_i and Eq. (2), $x_i(t)$ is calculated as follows,

$$x_i(t) = D_{im}\frac{d - r}{d_i}s\left(t - \frac{d_i}{c}\right), \tag{14}$$

where c (=340 m/s) is sound velocity. D_{im} is directivity of shotgun microphones [14], as shown in the following equation,

$$D_{im} = \begin{cases}\cos\theta_{im} & (|\theta_{im}| \le 90^\circ)\\ 0 & (|\theta_{im}| > 90^\circ)\end{cases}, \tag{15}$$

where θ_{im} is the angle of incidence of the sound source in the ith microphone, as shown in Fig. 2. Thus, a_i and t_i in Eq. (1) are denoted as follows,

$$a_i = \frac{d - r}{d_i}, \quad t_i = \frac{d_i}{c}. \tag{16}$$

3.2 Measurement of Head-Related Impulse Response

The amplitude and the initial delay of HRIRs ($a_{iL(iR)}$ and $t_{iL(iR)}$) in Eq. (3) must be estimated from measured HRIRs to construct the localization model. The following is the procedure for measuring HRIRs.

In a low-reverberant room, 24 loudspeakers (Emic: Soundevice) were placed in the circle with a radius of 2 m at intervals of 15° and a head and torso simulator (HATS) was placed in the center of the circle. The ears of the HATS were at the same height as the loudspeakers. The azimuth angle of the loudspeakers were -165°, -150°, ..., 0°, 15°, ..., 165°, and 180°. Note that 0° is the azimuth angle of the front direction of HATS. HRIRs were measured by playing a time stretched pulse (TSP) signal [15] from each loudspeaker. The sampling frequency and duration of the TSP were 48 kHz and 65536 samples, respectively. To reduce the room reverberation effect, measured HRIRs in the initial 440 samples, where the direct sound from loudspeakers comes only to HATS, were truncated.

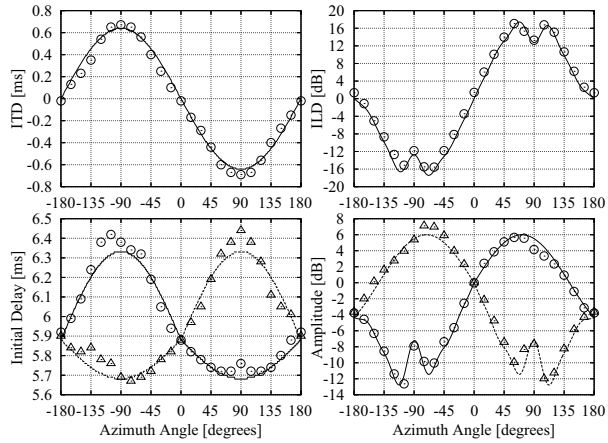


Fig. 3 Calculated and estimated HRIR parameter results.

3.3 Estimation of ITD and Initial Delay

Since the ITD contributes to the directional perception of sound sources at frequencies of less than 1.6 kHz [4], ITDs and initial delays are estimated from measured HRIRs, which are processed by a low-pass filter of 1.6 kHz. ITD $\Delta t_{LR}(= t_{iL} - t_{iR})$ is the peak time of the inter-aural correlation function calculated from measured HRIRs, as shown in the following equation,

$$\Delta t_{LR} = \arg\max_{\tau} \{h'_{iL}(t) h'_{iR}(t-\tau)\}, \tag{17}$$

where $h'_{iL}(t)$ and $h'_{iR}(t)$ are measured HRIRs from the loudspeaker of θ_i azimuth angle to both ears. Calculated results of ITD Δt_{LR} are shown in the upper left of Fig. 3 at points on the circle. ITDs Δt_{LR} are estimated from the calculated results as follows,

$$\Delta t_{LR}(\theta_i)[\text{ms}] = -0.64 \sin\theta_i. \tag{18}$$

Estimated results are shown as the solid line in the upper left panel of Fig. 3. We think that the ITDs are estimated with satisfactory accuracy because the mean square error between calculated results and estimated results is 0.046 ms.

Initial delays (t_{iL} and t_{iR}) are the times when the initial peak comes in the waveform of measured HRIRs, which are processed by the low-pass filter. Calculated results for t_{iL} and t_{iR} are shown in the lower left panel of Fig. 3 as circles and triangles, respectively. The t_{iL} and t_{iR} are estimated from calculated results as follows,

$$\begin{aligned} t_{iL}(d'_i, \theta_i)[\text{ms}] &= \begin{cases} \frac{1000 d'_i}{c} - 0.16 \sin\theta_i \, (\theta_i \geq 0^\circ) \\ \frac{1000 d'_i}{c} + 0.16 \sin\theta_i \\ \qquad + \Delta t_{LR}(\theta_i) \, (\theta_i < 0^\circ) \end{cases}, \\ t_{iR}(d'_i, \theta_i)[\text{ms}] &= t_{iL}(d'_i, -\theta_i), \end{aligned} \tag{19}$$

where d'_i is calculated from Fig. 2 as follows,

$$d'_i = \sqrt{(r\cos\theta_i - \Delta_x)^2 + (r\sin\theta_i - \Delta_y)^2}. \tag{20}$$

Estimated results are shown as the solid and dashed lines in the lower left panel of Fig. 3. We think that initial delays are estimated with satisfactory accuracy because the mean square error between calculated results and estimated results is 0.043 ms.

3.4 Estimation of ILD and Amplitude

Since the inter-aural level difference (ILD) contributes to the perceived direction of sound sources in full band frequencies [4], ILDs and amplitudes are estimated from measured HRIRs that are not processed by filters. Amplitude of the ears (a_{iL} and a_{iR}) is the average power of measured HRIRs of the ears. The unit of amplitude is dB and refers to the amplitude at 0° direction. ILDs $\Delta a_{LR}(= a_{iL}/a_{iR})$ were calculated from amplitudes of both ears. Calculated results are shown in the upper right panel of Fig. 3 as circles. ILDs are estimated from calculated results as follows,

$$\Delta a_{LR}(\theta_i)[\text{dB}] = \sum_{k=1}^{7} Q_k \sin q_k \theta_i, \tag{21}$$

where q_k and Q_k are parameters denoted as follows,

$$\begin{aligned} \{q_k\} &= 1, 2, 4, 5, 7, 9, 11, \\ \{Q_k\} &= 16.06, 1.64, 0.70, -1.36, 0.88, -0.70, 0.37. \end{aligned} \tag{22}$$

Estimated results are shown in the upper right panel of Fig. 3 at the solid line. We think that the ILDs are estimated with satisfactory accuracy because the mean square error between calculated results and estimated results is 1.035 dB.

Calculated results for a_{iL} and a_{iR} are shown in the lower right panel of Fig. 3 as circles and triangles, respectively. The a_{iL} and a_{iR} are estimated from the calculated results as follows,

$$\begin{aligned} a_{iL}(d'_i, \theta_i)[\text{dB}] &= \begin{cases} 20 \log_{10} \frac{2}{d'_i} + 5.99 \sin(\frac{180^\circ}{143^\circ}\theta_i) \\ (\theta_i \geq 0^\circ) \\ 20 \log_{10} \frac{2}{d'_i} - 5.99 \sin(\frac{180^\circ}{143^\circ}\theta_i) \\ +\Delta a_{LR}(\theta_i) \, (\theta_i < 0^\circ) \end{cases}, \\ a_{iR}(d'_i, \theta_i)[\text{dB}] &= a_{iL}(d'_i, -\theta_i). \end{aligned} \tag{23}$$

Estimated results are shown in the lower right panel of Fig. 3 as the solid and dashed lines. We think that amplitudes are estimated with satisfactory accuracy because the mean square error between calculated and estimated results is 0.676 dB.

3.5 Prediction of Direction

From Eq. (18), the relation between ITDs τ and perceived directions ϕ is derived as follows,

$$\tau = -0.64 \sin\phi \qquad (-90^\circ \leq \phi \leq 90^\circ). \tag{24}$$

Therefore, by applying Eq. (24) to Eq. (11), sound image direction ϕ' is predicted as follows,

$$\phi' = \sin^{-1}\left(-\frac{\tau_E}{0.64}\right), \tag{25}$$

where the range of predicted directions is $-90° \leq \phi' \leq 90°$. If $-\frac{\tau_E}{0.64} < -1$ and $-\frac{\tau_E}{0.64} > 1$, the predicted direction is $\phi' = -90°$ and $90°$.

3.6 Precedence Effect Coefficient

Effective ITDs τ_E in Eq. (25) were calculated from parameters obtained in section 3.1 (a_i, t_i, and D_{im}), section 3.3 (t_{iL} and t_{iR}), section 3.4 (a_{iL} and a_{iR}), and precedence effect coefficient (p_i). D_{is} was the directivity of the omnidirectional loudspeakers ($D_{is} = 1$) [14].

p_i in Eq. (6) was calculated from t_i and $t'_i (= d'_i/c)$ obtained in section 3.1. The value of α in Eq. (6) was set according to following equation,

$$\alpha = \arg\min_{\alpha} \sqrt{\frac{\sum_{\phi}(\phi' - \phi_0)^2}{L \times S \times C \times K}}, \tag{26}$$

where ϕ' and ϕ_0 are the predicted results of the localization model and the localized results of the listening tests [6], [7] in the presented direction ϕ. $L(= 7)$, $S(= 2)$, $C(= 5)$, and $K(= 4)$ are the total number of presented directions, dry sources, conditions related to the number of channel signals, and listening positions in the listening tests. As a result, α was set to 5.25×10^3. In the lower limit time of the precedence effect (between 0.63 ms and 1 ms [4]), the value of p_i is between 0.0052 and 0.0366 according to following calculations,

$$\begin{aligned} p_i &= \exp\{5.25 \times 10^3 \times (-1 \times 10^{-3})\} = 0.0052 \\ &\qquad (t_i + t'_i - t_{\min} = 1\,\text{ms}), \\ p_i &= \exp\{5.25 \times 10^3 \times (-0.63 \times 10^{-3})\} = 0.0366 \\ &\qquad (t_i + t'_i - t_{\min} = 0.63\,\text{ms}). \end{aligned} \tag{27}$$

We think that the reduction of the contribution to the directional perception by the precedence effect coefficient p_i is adequately expressed.

4. Comparison with Listening Test

To evaluate the effectiveness of the proposed model described in Section 3, the results predicted by the proposed model were compared with the localized results of the listening tests [6], [7].

4.1 Listening Test Procedure [6], [7]

The listening test was done in a low-reverberation room with a reverberation time of about 80 ms. Twenty-three loudspeakers (Emic: Soundevice) were placed at the front of the listening area, which was a circle with a radius of two meters, and four listening positions were placed as shown in Figs. 4 and 5. The gray circles indicate sound images reproduced by the loudspeaker array. The sound played from loudspeakers was synthesized by Eq. (14). The sampling

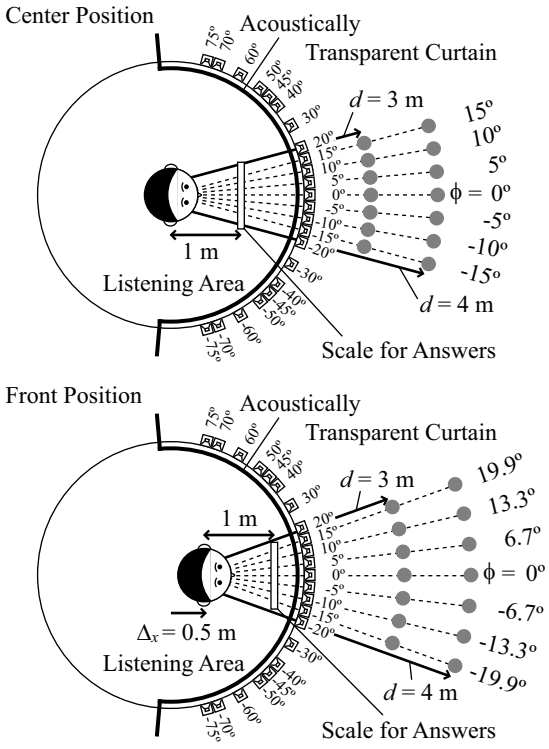


Fig. 4 Environment in listening test (Center and front positions) [7].

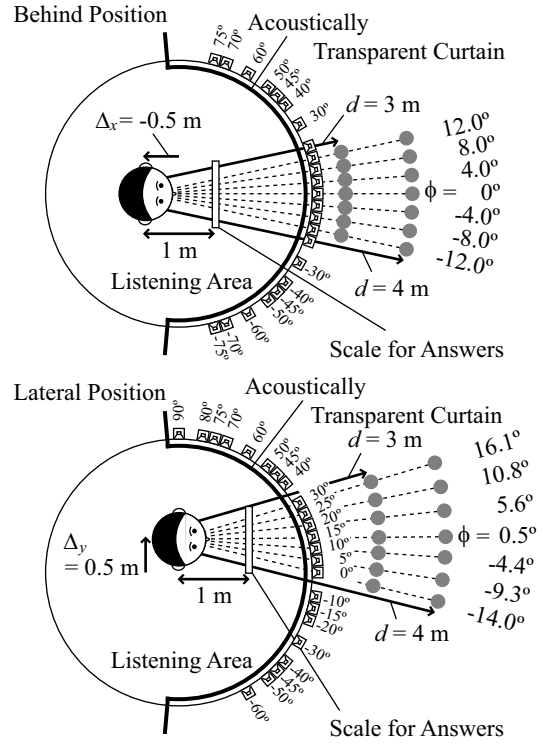


Fig. 5 Environment in listening test (Behind and lateral positions) [7].

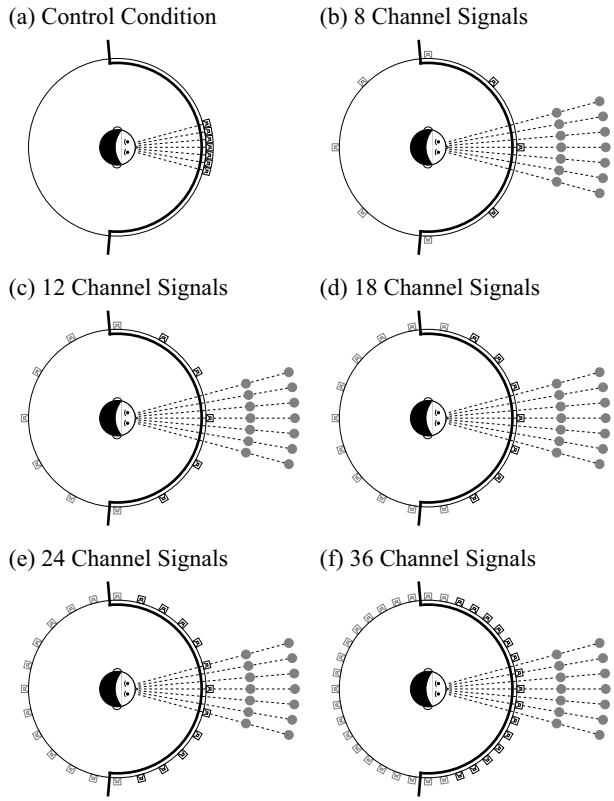


Fig. 6 Experimental conditions (number of channel signals in listening test) (Center position).

frequency of the sound is 48 kHz. Since the channel signals at the back of the listening area always become zero when sound images are placed at the front of the listening area, no loudspeakers were placed at the back of the area. The level of background room noise was 25.0 dB(A), and the sound pressure level was set at about 70 dB(A) at the center of the circle. The subjects were unable to see the loudspeakers because they were hidden behind an acoustically transparent curtain.

The five conditions, which correspond to the number of channel signals, are shown in Fig. 6. The loudspeakers drawn in gray indicate loudspeakers that are not placed according to the directivity of microphones. In the control condition (a), the sound source itself was presented to subjects by playing a dry source from one loudspeaker selected from a group of seven. Since the azimuth angle of the sound source varies in the control condition when the listening position is not the center position, the value of the azimuth angle is calculated based on the listening position, as shown in Eq. (28),

$$\phi_{\text{front,behind,lateral}} = \tan^{-1}\left[\frac{r\sin\phi_{\text{center}} - \Delta_y}{r\cos\phi_{\text{center}} - \Delta_x}\right], \tag{28}$$

where ϕ_{center}, ϕ_{front}, ϕ_{behind} and ϕ_{lateral} is the azimuth angle of the sound source in the center, front, behind, and lateral positions. In this test, the values of the azimuth angle are denoted as follows,

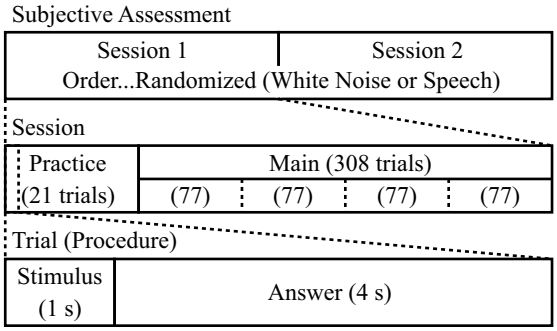


Fig. 7 Flowchart of listening test [7].

Table 1 Trial conditions for listening tests [7].

	Element	Note
Practice (21)	= 7 directions ×(1 condition × 2 distances + control)	 (f) of Fig. 6 3, 4 m (a) of Fig. 6
Main (308)	= 7 directions ×(5 conditions × 2 distances + control) × 4 repetitions	 (b)–(f) of Fig. 6 3, 4 m (a) of Fig. 6

$$\begin{aligned}
\phi_{\text{center}} &= -15, -10, -5, 0, 5, 10, 15^\circ, \\
\phi_{\text{front}} &= -19.9, -13, 3, -6.7, 0, 6.7, 13.3, 19.9^\circ, \\
\phi_{\text{behind}} &= -12.0, -8.0, -4.0, 0, 4.0, 8.0, 12.0^\circ, \\
\phi_{\text{lateral}} &= -14.0, -9.3, -4.4, 0.5, 5.6, 10.8, 16.1^\circ,
\end{aligned} \tag{29}$$

where ϕ_{center}=0, 5, 10, 15, 20, 25, 30° in the lateral position. In conditions (b)–(f), channel signals were played from three, five, seven, eleven, and fifteen loudspeakers chosen from twenty-three loudspeakers. A zero signal was assigned to the loudspeakers that were not chosen. Subjects reported feeling that there are synthetic sound images in the positions occupied by the gray circles, as shown in Fig. 6.

Subjects were twelve students (ten males and two females). Three subjects were placed in each listening position. The experimental design of the listening test is shown in Fig. 7. The test was divided into two sessions for each dry source (white noise and speech). The order of presentation of the dry sources was randomized for each subject. In each session, after 21 practice trials, 308 main trials were done. Rest periods were allowed after every 77 main trials. The conditions for the practice and main trials are shown in Table 1. The subject was instructed to report the direction of the sound within four seconds after listening to a one-second stimulus. Subjects reported the direction on a scale that was placed one meter in front of the listening position, as shown in Figs. 4 and 5. This scale is marked from -25° to 25° at 2.5° intervals. The subjects can turn their heads freely during listening tests.

Localized results in the control condition are shown in Fig. 8. In all listening positions, perceived directions are about the same as presented directions. The mean square error between presented and perceived directions is 1.817°.

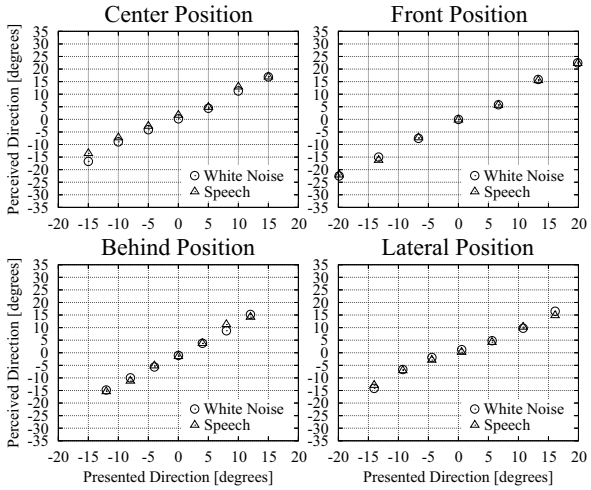


Fig. 8 Results of listening test (Control condition).

Thus, we think that it is possible to compare the predicted results of the proposed model with the localized results of the listening test because the subjects can accurately localize the direction of sound sources.

4.2 Comparison of Results

The localized direction of the proposed model was predicted as shown in Eq. (25). The presented direction ϕ was from -20° to 20°. The localized direction of the conventional model was also predicted. In the conventional model, as in Kurozumi et al.'s model [12], $\alpha = 0$ and $p_i = 1$ in Eq. (6).

The predicted direction ϕ' of the conventional model and the proposed model and the perceived direction of the listening test are shown in Figs. 9–13. When there were 8, 12, and 18 channel signals, the perceived direction was not the same as the presented direction. This is due to the increase in the precedence effect [6], [7].

Since the precedence effect is not introduced in the conventional model, the predicted direction of sound images differs greatly from the perceived direction when the perceived direction is biased and the listening position is the lateral position. However, since the precedence effect is introduced in the proposed model, the predicted direction of sound images is about the same as the perceived direction, even if the perceived direction is biased and the listening position is the lateral position. This shows that the precedence effect must be used in the localization model for the localized direction in the sound field reproduction system based on wave field synthesis to be accurately predicted.

For our quantitative evaluation, mean square errors (MSEs) between the perceived direction and the predicted direction were calculated as shown in Eq. (30),

$$\text{MSE [degrees]} = \sqrt{\frac{\sum_{\phi}(\phi' - \phi_0)^2}{L \times S}}, \tag{30}$$

where ϕ' and ϕ_0 are the predicted and perceived directions

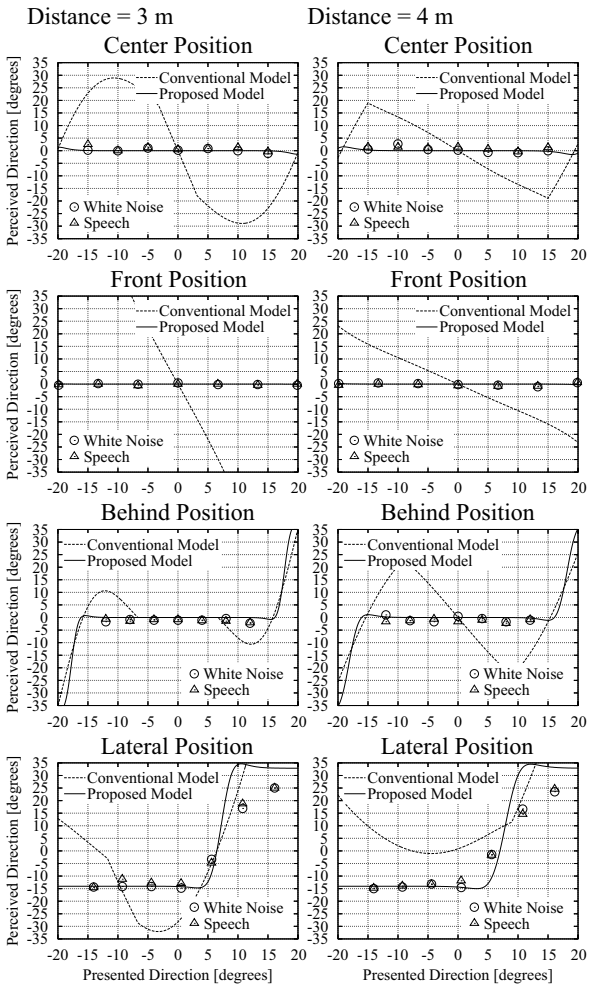


Fig. 9 Results of listening test and localization model (8 channel signals).

in the presented direction ϕ, $L(=7)$ and $S(=2)$ are the total number of presented directions and dry sources. The results of the MSEs calculated in each model (conventional and proposed) and each source distance (3, 4 m) are shown in Fig. 14. The MSEs in the conventional model are large on the whole. This is especially when the MSEs are more than 10° and when there are eight channel signals and the listening position is the lateral position. In contrast, there are fewer MSEs in the proposed model than in the conventional model when there are eight channel signals and the listening position is the lateral position. The MSEs in the proposed model are less than 5°. Since this value is greater than that of the difference limen of a broadband noise source in the front direction (about 3° [4]), the accuracy of the proposed model is not adequate from the point of view of the auditory system. However, this value is less than that of the difference limen in the front direction in the ventriloquism effect (at least 11° [16]), meaning that the proposed model is can accurately predict the localized results of listening tests, thereby reducing the cost of listening tests in the construction of an audio-visual virtual reality system.

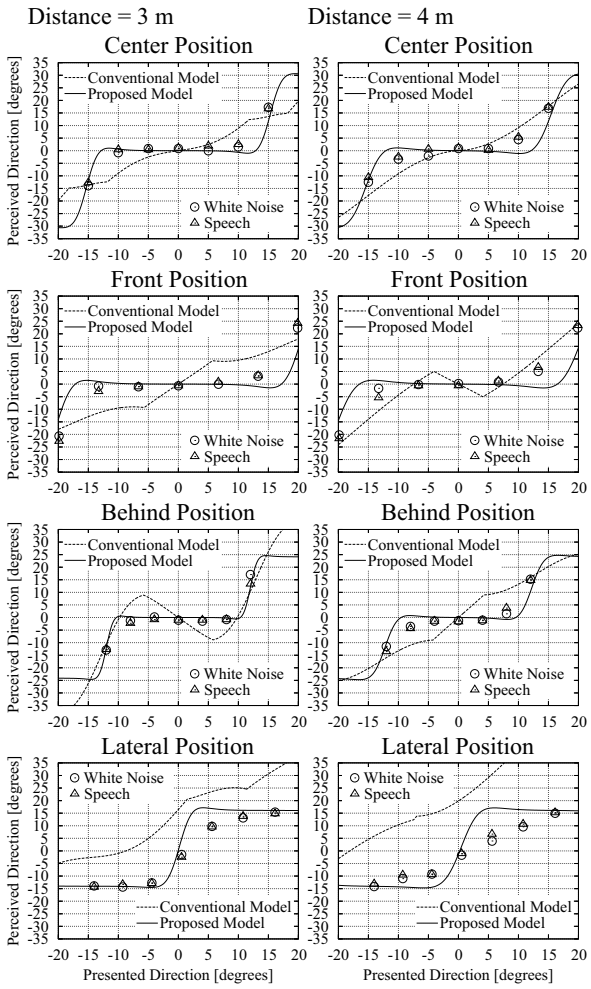


Fig. 10 Results of listening test and localization model (12 channel signals).

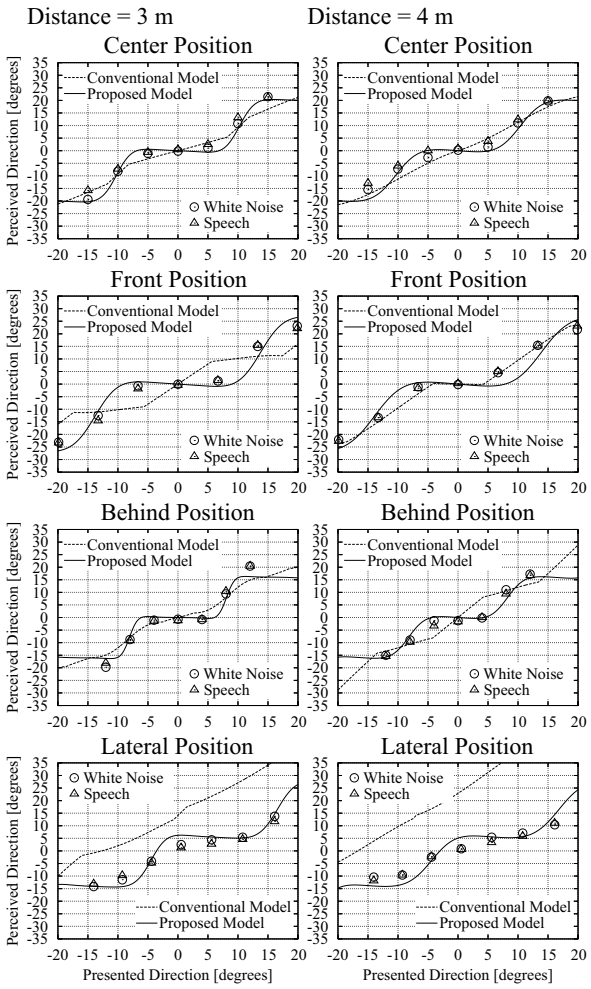


Fig. 11 Results of listening test and localization model (18 channel signals).

5. Conclusion

We proposed a localization model of synthesized sound images that predicts the results of listening tests. This model will reduce the costs of listening tests in sound field reproduction based on wave field synthesis. In the proposed model, the precedence effect is introduced and the direction of synthesized sound images is predicted based on inter-aural time differences. Our comparison of the predicted results of the proposed model and the localized results of the listening test shows that the proposed model can accurately predict the localized results of listening tests.

Since the model proposed in this paper is based on the inter-aural time difference, the predicted direction is limited to the front direction of the horizontal plane. However, in realistic systems, sound images are presented from the front and behind and upside directions. Localization models that can predict the localized results of listening tests where sound images are presented from the behind or upside direction need more study.

Acknowkedgement

The authors would like to thank Prof. K. Takeda, Prof. F. Itakura and Prof. M. Nakagawa for their continued support and encouragement during our research.

References

[1] H. Fletcher, “Symposium on wire transmission of symphonic music and its reproduction on auditory perspective: Basic requirement,” Bell System Technical Journal, 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] J. Blauert, Spatial Hearing, revised ed., MIT Press, Cambridge, Mass, 1997.

[5] M.R. Schroeder, D. Gottlob, and K.F. Siebrasse, “Comparative study of european concert halls: Correlation of subjective preference with geometric and acoustic parameters,” J. Acoust. Soc. Am., vol.56, no.4, pp.1195–1201, October 1974.

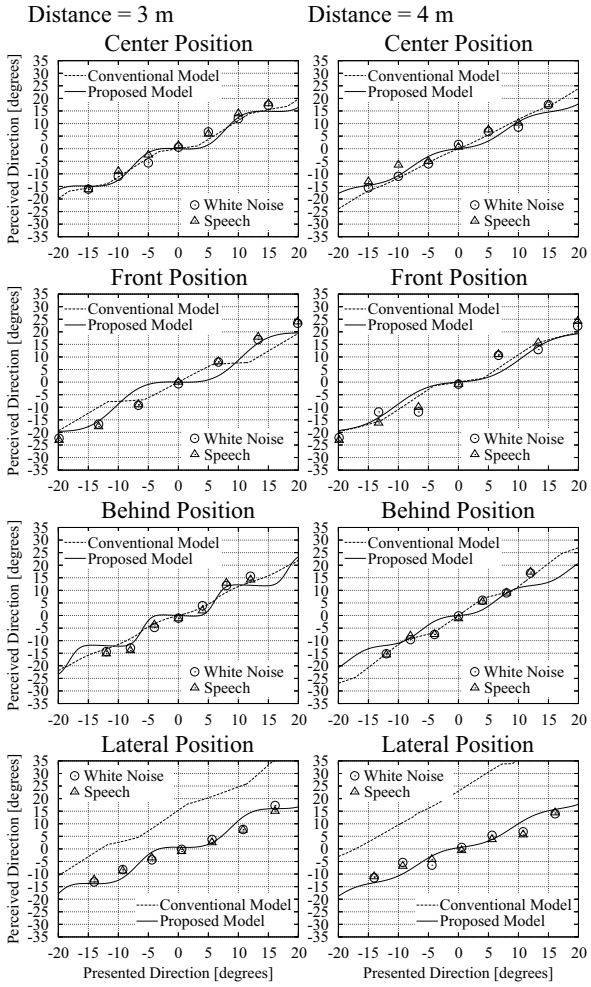


Fig. 12 Results of listening test and localization model (24 channel signals).

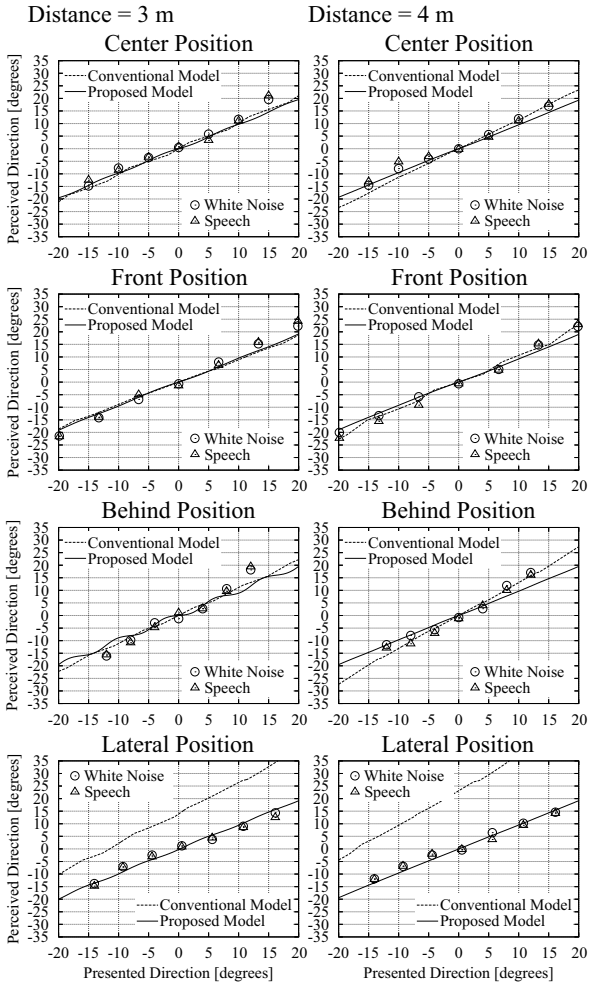


Fig. 13 Results of listening test and localization model (36 channel signals).

[6] T. Kimura, K. Kakehi, K. Takeda, and F. Itakura, "Subjective effect of the number of channel signals in wave field synthesis - in the case of sound sources of frontal direction -," Trans. VR Soc. Jpn., vol.10, no.2, pp.257–266, June 2005. (in Japanese).

[7] T. Kimura, A Study of the Method for the Reduction of Information in Sound Field Reproduction Based on Wave Field Synthesis, thesis for the Ph.D. degrees, Graduate School of Human Informatics, Nagoya University, December 2005. (in Japanese).

[8] M. Morimoto, "The relation between spatial impression and the precedence effect," Proceedings of International Conference on Auditory Display, Kyoto, Japan, no.SS#2-1, pp.297–306, July 2002.

[9] L.A. Jeffress, "A place theory of sound localization," J. Comp. Physiol. Psychol., vol.61, pp.468–486, 1948.

[10] W. Lindemann, "Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation of lateralization for stationary signals," J. Acoust. Soc. Am., vol.80, no.6, pp.1608–1622, December 1986.

[11] W. Lindemann, "Extension of a binaural cross-correlation model by contralateral inhibition. II. The law of the first wave front," J. Acoust. Soc. Am., vol.80, no.6, pp.1623–1630, December 1986.

[12] K. Kurozumi and K. Ohgushi, "A model for sound localization based on the interaural-correlation function," J. Acoust. Soc. Jpn., vol.44, no.10, pp.726–734, October 1988. (in Japanese).

[13] H. Wallach, E.B. Newman, and M.R. Rosenzweig, "The precedence effect in sound localization," Am. J. Psycho., vol.62, no.3, pp.315–336, July 1949.

[14] T. Kimura and K. Kakehi, "Effects of directivity of microphones and loudspeakers on accuracy of synthesized wave fronts in sound field reproduction based on wave field synthesis," Papers of AES 13th Regional Convention, Tokyo, Japan, no.0037, pp.1–8, July 2007.

[15] Y. Suzuki, F. Asano, H.Y. Kim, and T. Sone, "An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses," J. Acoust. Soc. Am., vol.97, no.2, pp.1119–1123, February 1995.

[16] S. Komiyama, "Interaction between vision and auditory in spatial perception," J. Acoust. Soc. Jpn., vol.52, no.1, pp.46–50, January 1996. (in Japanese).

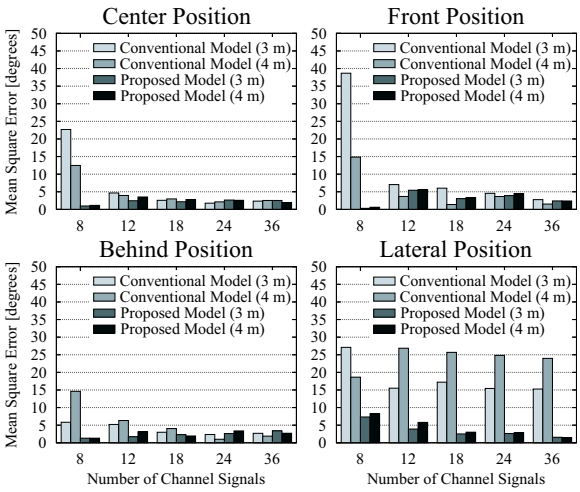
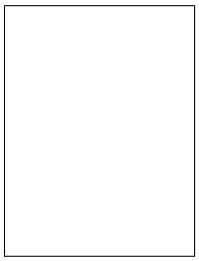
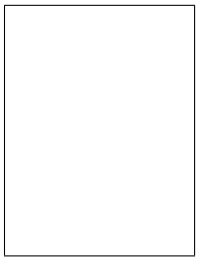


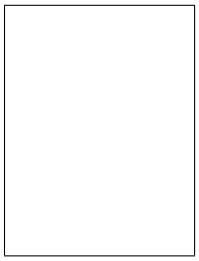
Fig. 14 Mean square errors of conventional and proposed model.



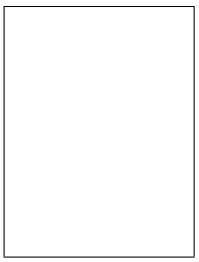
Toshiyuki Kimura received the B.E., M.A. and Ph.D. degrees in Nagoya University, Japan in 1998, 2000 and 2005, respectively. He was a research fellow of Japan Society for the Promotion of Science, a research fellow of Nagoya University, Japan, and a research associate of Tokyo University of Agriculture and Technology, Japan from 2003 to 2007. He is currently an expert researcher of National Institute of Information and Communications Technology, Japan from 2007. His research interests include sound field reproduction, array signal processing and spatial hearing. He is a member of ASJ, VRSJ and AES.



Yoko Yamakata received the B.S. degree in Engineering, M.S. and Ph.D. degrees in Informatics from Kyoto University, Japan, in 2000, 2002 and 2006, respectively. She was a research associate of Graduate School of Informatics, Kyoto University from 2005 to 2006. She is currently an expert researcher of National Institute of Information and Communications Technology, Japan from 2006. Her research interests include multi-media processing, 3D sound field reproduction, and audio transducer. She is a member of JSAI, ASJ and ITE.



Michiaki Katsumoto



Kazuhiko Kakehi received the B.S., M.E. and D.E. degrees from Waseda University, Japan in 1965, 1967 and 1993, respectively. He joined NTT Musahino Labs. in 1967, where he mainly studied about speech quality in telecommunication and hearings. He moved to the Graduate School of Human Informatics at Nagoya University, Japan as a professor in 1994. Since 2004, he is a professor of School of Information Science and Technology at Chukyo University, Japan. His main research interest is human sentence and speech processing, human interface, and hearings. He is a member of ASJ, JCSS, PSJ, JSHBD and LSJ.