I. Nambu, M. Washizu, S. Morioka, Y. Hasegawa, W. Sakuma, S. Yano, H. Hokari and Y. wada,
"Reinforcement-Learning-Based Personalization of Head-Related Transfer Functions,"
J. Audio Eng. Soc., vol. 66, no. 5, pp. 317–328, (2018 May.).
DOI: https://doi.org/10.17743/jaes.2018.0014

Reinforcement-Learning-Based Personalization of Head-Related Transfer Functions

ISAO NAMBU[1], MANABU WASHIZU[1], SHUHEI MORIOKA[1], YUTA HASEGAWA[1], WATARU SAKUMA[1], SHOHEI YANO[2], HARUHIDE HOKARI[1], AND YASUHIRO WADA[1]

[1]Nagaoka University of Technology, Nagaoka, Japan
[2]National Institute of Technology, Nagaoka College, Nagaoka, Japan

Accurately perceiving spatial locations of virtual sounds using stereo earphones or headphones requires individual head-related transfer functions (HRTFs) for each listener. However, accurate HRTF measurement is usually difficult. While previous studies have proposed methods of HRTF personalization without HRTF measurement, localization errors often remain and further modifications are challenging. We therefore propose a method that uses reinforcement learning and listener evaluation to obtain an accurate individual HRTF without measurement. We conducted a proof-of-concept simulation and an experiment involving human subjects. In the simulation, we confirmed that the proposed method could acquire individual HRTFs close to the measured HRTF from a dummy-head HRTF. Next, we conducted a learning experiment in one direction using the proposed method without individual HRTFs and observed improved horizontal-plane localization for the learned HRTF compared to the dummy-head HRTF. These results collectively demonstrate the possibility of the proposed reinforcement-learning-based personalization method for individual HRTFs that enables listeners to experience accurate virtual sound environments.

0 INTRODUCTION

Virtual sound synthesis is a method for auditory images that are perceived at spatial locations outside the head with stereo earphones or headphones [1–3]. This method is achieved by equalizing sound from loudspeakers and earphones, respectively, to the eardrum. To generate accurate virtual sound, a head-related transfer function (HRTF) is needed. The HRTF is a transfer function that represents the characteristics of a direction-dependent transformation into a sound signal, which is imposed by the head and pinna of the listener. By filtering arbitrary sound through an HRTF, one can produce virtual sound using earphones. Because the shapes of the head and pinna are different for each individual [2, 3], individual HRTFs are required for accurate sound localization. However, measuring an individual HRTF is impractical because it requires a special environment such as a large anechoic chamber.

Several methods have been proposed to obtain HRTF without measurement. One approach is to use anthropometric features of the listener's ear, head, and torso [4–8]. Based on the similarity of the anthropometric features, these methods select one of the HRTFs in the database. Additionally, some recent advanced works have shown that the morphological information obtained via magnetic resonance imaging can be used to develop an acoustic simulator and to generate a simulated HRTF based on numerical analysis techniques [9, 10]. These methods can be the morphological information very accurately but requires special environments.

Another approach is to select a well-localized HRTF from candidates using a listener's evaluation [11–13]. For example, Middlebrooks et al. [11] and Fink and Ray [13] analyzed HRTFs using principal component analysis (PCA), and they achieved better localization by changing the principal component weights based on a listener's evaluation and no special environment is necessary.

However, even when using these methods, localization errors often remain because many conventional methods do not estimate the user's own HRTF directly [4–7, 11–13] because these methods can be used to select good (or the best) HRTFs from the prepared database. Therefore, further improvement is difficult even when the selected "best" HRTF does not provide good localization accuracy.

To overcome this issue, we propose a method based on reinforcement learning to modify the HRTF according to the user's feedback. Reinforcement learning is an algorithm for learning a function in which the maximum reward (user's evaluation in this case) is achieved without a supervised signal (i.e., information of the measured HRTFs in this case)

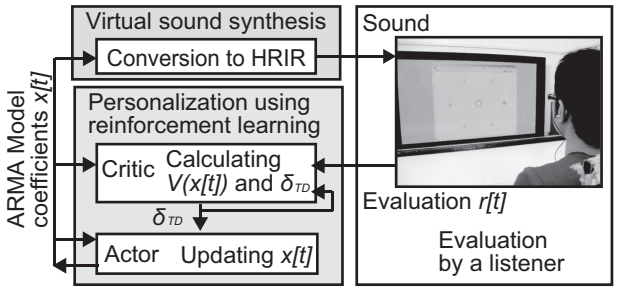


Fig. 1. Overview of the proposed method.

[14, 15] (see below or Appendix 1 for details). Our method does not have any restrictions or assumptions regarding the shape of the HRTF. Therefore, we expect that it can estimate any HRTF from the subjective evaluation alone, even when information on how to modify the HRTF is unavailable.

We herein present a simulation and an experiment to acquire an individual HRTF from the dummy-head HRTF as a proof-of-concept of our proposed method. Thus, by setting the reward as the user's evaluation for sound localization, we can obtain an individual HRTF that provides good localization without an HRTF measurement.

1 REINFORCEMENT-LEARNING-BASED HRTF PERSONALIZATION METHOD

We propose a method to obtain an accurate individual HRTF from an existing one using reinforcement learning. Based on a user's evaluation for sound localization, this method searches for parameters for the HRTF in order to acquire an accurate individual HRTF (Fig. 1). Our proposed method consists of two parts: HRTF modeling and reinforcement learning.

1.1 Modeling of HRTF

In the present study we define HRTF as a transfer function of the input sound signals of loudspeakers into the output signals of the microphone placed in the ear canal (sometimes referred to as a spatial sound transfer function). This HRTF includes the head and pinna characteristics as well as the loudspeakers, microphones at the ear canal, and room reverberation. Note that this definition differs from those described in other works (e.g., [1]). Virtual sounds are created as signals convolved with a head-related impulse response (HRIR), which is the impulse response corresponding to an HRTF, through the stereo earphones. This procedure for generating virtual sounds is independently performed for each ear.

In the modeling process the HRIR for the dummy-head is modeled by an auto-regressive moving average (ARMA) model [16] to reduce the number of parameters because the number of HRIR samples is too large (128 in this study) to use reinforcement learning. Before modeling, the initial time delay was removed using the method proposed by Nishino et al. [7]. This delay was recovered when the virtual sound was generated. After this processing, the ARMA modeled impulse response at the k-th sample $\hat{h}[k]$ was

$$\hat{h}\ [k] = \sum_{n=1}^{P} p\,[n]\,\hat{h}\,[k-n] + \sum_{m\,=\,0}^{Q} q\,[m]\,\delta\,[k-m] \tag{1}$$

where p and q are AR and MA parameters, and P and Q are the number of AR and MA parameters, respectively. δ is the Dirac delta function. p and q were each calculated using the least square method. P and Q were determined in order to minimize the cost function J, which is the difference between the measured HRIR h and $\hat{h}$ [16],

$$J\ = \frac{\sum_{k=0}^{N}\left(h\,[k]-\hat{h}\,[k]\right)^2}{\sum_{k=0}^{N} h^2\,[k]}\ , \tag{2}$$

where N is the number of samples for HRIR. This modeling results in a reduced number of parameters for reinforcement learning.

1.2 Reinforcement-Learning-Based Personalization

The proposed method is additionally comprised of reinforcement learning (Fig. 1). This algorithm is used to modify the HRTF using the reward r provided by the user in a sound-localization task. We use the actor-critic reinforcement learning method with continuous state modeling for discrete time [14, 15], which enables fine-tuning of the HRTF. At a trial iteration t, the actor selects the appropriate coefficients of ARMA models $x[t]$ as outputs. The critic evaluates the values of the selected coefficients $V(x[t])$ and calculates the temporal-difference (TD) error δ_{TD} based on the reward at trial t ($r[t]$). This error is used for updating the actor and critic [14, 15]. For further details of the algorithm and parameters, refer to Appendix 1 and [14, 15].

2 SIMULATION

The objective of the simulation was to examine whether the proposed method can learn the target HRTF from the dummy-head HRTF. Assuming that the dummy-head HRTF has frequency ranges that are required for sound localization for human subjects, we used this HRTF as an initial value and performed simulations to obtain an individual HRTF.

2.1 Simulation Setting

This study was approved by the Ethics Committee of Nagaoka University of Technology. All human subjects were given instructions about the study and provided informed consent before the experiment. We measured the HRIRs for a dummy-head (SAMRAI, Koken, Japan) and for nine men (ages 20 to 24) in a sound-proof room (reverberant times: 87 ms) at the Sound Vibration Engineering Center of the Nagaoka University of Technology. Sixteen-ordered M-sequence signals were generated from loudspeakers (SD-0.6, Soundevice, Japan). The impulse

responses were measured by miniature electret condenser microphones (UC-92, Rion, Japan) placed at the entrance of the external ear canal. Loudspeakers were located along 24 directions at 15° intervals in the horizontal plane with 1.5 m distance. We defined the direction of each loudspeaker in front of the subject or dummy-head as 0°. The right side of the 0° loudspeaker had positive angles whereas the left side had negative angles. Considering the convergence of HRIR, the sampling number used for modelling was 128 with a frequency of 44.1 kHz.

The simulation was conducted using MATLAB (Mathworks Inc., USA). HRTFs for 24 directions were tested for each subject. A set comprised of an HRTF modification and an evaluation was defined as a trial. In a single simulation, 300,000 trials were conducted for each direction. We conducted this simulation 30 times with randomly changing initial search noise. Personalization was said to succeed when the reward was reduced in comparison to the initial condition.

To simplify the setting of parameters, each ARMA parameter ($x[t]$) was normalized so that its target value was within ± 1 for each direction for each subject. For the actor and critic, the number of intermediate nodes was initially set to 1. The averaged number of the ARMA parameters (P and Q in the Eq. (1)) was 8.8 (standard deviation $\pm$ 8.1) and 54.6 ($\pm$ 8.1), respectively.

2.2 Reward Setting

To determine whether our method can generate an individual HRTF in which virtual sound is accurately localized, the measured individual HRTFs for each subject were used as the target HRTFs. Thus, we examined how our method modifies the dummy-head HRTF into the target HRTF using the reward. We used the spectral distortion (SD) [7] between the target and learned HRTFs as the reward r. This index indicates the differences between two HRTFs in the frequency domain, which is defined as

$$\begin{aligned} r &= -SD \\ &= -\sqrt{\frac{1}{N_f}\sum_{k=1}^{N_f}\left(20\log_{10}\frac{|H(\omega_k)|}{|\hat{H}(\omega_k)|}\right)^2}\,[\mathrm{dB}]\,. \end{aligned} \tag{3}$$

Here, $H(\omega_k)$ and $\hat{H}(\omega_k)$ are targets and learned HRTFs at frequency ω_k, respectively. N_f is the number of samples for HRTF. This value is 64 because HRTF was calculated from 128 samples of HRIR by the discrete Fourier transform. The HRTF was calculated in the frequency range from 500 Hz to 16 kHz because this range is the most-relevant for sound localization [17, 18]. Negative SD was set as a reward because a large SD indicates dissimilarity between the target and learned HRTFs. To determine whether our method can acquire actual localized HRTF, the measured individual HRTF for each subject was used as the target HRTF. (This target is not available when one wants to obtain an individual HRTF without its measurement.)

2.3 Simulation Results

For all 9 subjects and 24 directions, we performed 30 simulations and found that the learned HRTFs were close to the target HRTFs in almost all cases. An example of this improvement for the right 30° direction for three representative subjects is shown in Fig. 2(a). The target (individual) HRTF was obtained from the dummy-head HRTF. As shown in Fig. 2(b), the initial SD (reward) was influenced by the speed of convergence and required approximately 200,000 to 300,000 trials to obtain a stable SD at 1–2 dB. Similar results were obtained for all the subjects. On an average, 26.8 simulations ($\pm$1.7 standard deviation across subjects) out of 30 succeeded. This number corresponds to 89% of the total simulations. The SD when checked specifically after learning, showed a value less than 5 dB in 88% of the simulations and less than 2 dB in 77% of the simulations. The SD reached a value of less than 1 dB in about one-third of the simulations. At this level (1–2 dB), the user typically cannot differentiate between the HRTFs [19]. Even the initial SDs were greater than 10 dB, and decreasing SDs were confirmed for all 24 directions (Fig. 2(c)).

2.4 Evaluation of the Learned HRTF

In this study we used SDs to evaluate the learned HRTF. However, it has been suggested that SD does not fully describe the effect of the spectral cues on the sound localization [20, 21]. Thus, decreasing SDs might not be related to the actual localization. To check that the learned HRTFs after learning show better sound localization, we conducted the experiment involving human subjects.

2.4.1 Protocol

Of the subjects, for which HRTFs were measured in the simulation, six out of nine participated in this experiment. HRTFs for both ears were prepared in order to generate virtual sounds. We examined two evaluations; the localization accuracy and subjective evaluation (five-grade evaluation) for the measured HRTF, learned HRTF after the simulation (Sec. 2.3), and dummy-head HRTFs.

In the localization accuracy evaluation, the subjects were asked to immediately determine the direction of the sound and select 1 out of the 12 directions (forced choice). In the five-grade evaluation, the subjects were asked to assess the extent to which the direction of the evaluation sound deviated from that of the reference sound using five criteria/grade. The detailed protocol is described in Appendix 2.

2.4.2 Evaluation Results

Results of the localization accuracies and the five-grade evaluation are shown in Fig. 3. The learned and measured HRTFs showed similar localization results. However, the results from the dummy-head HRTF differed. This similarity and difference is easily observed in the five-grade evaluation. Similar to the measured HRTF, the learned HRTF is rated almost five, meaning that the perceived

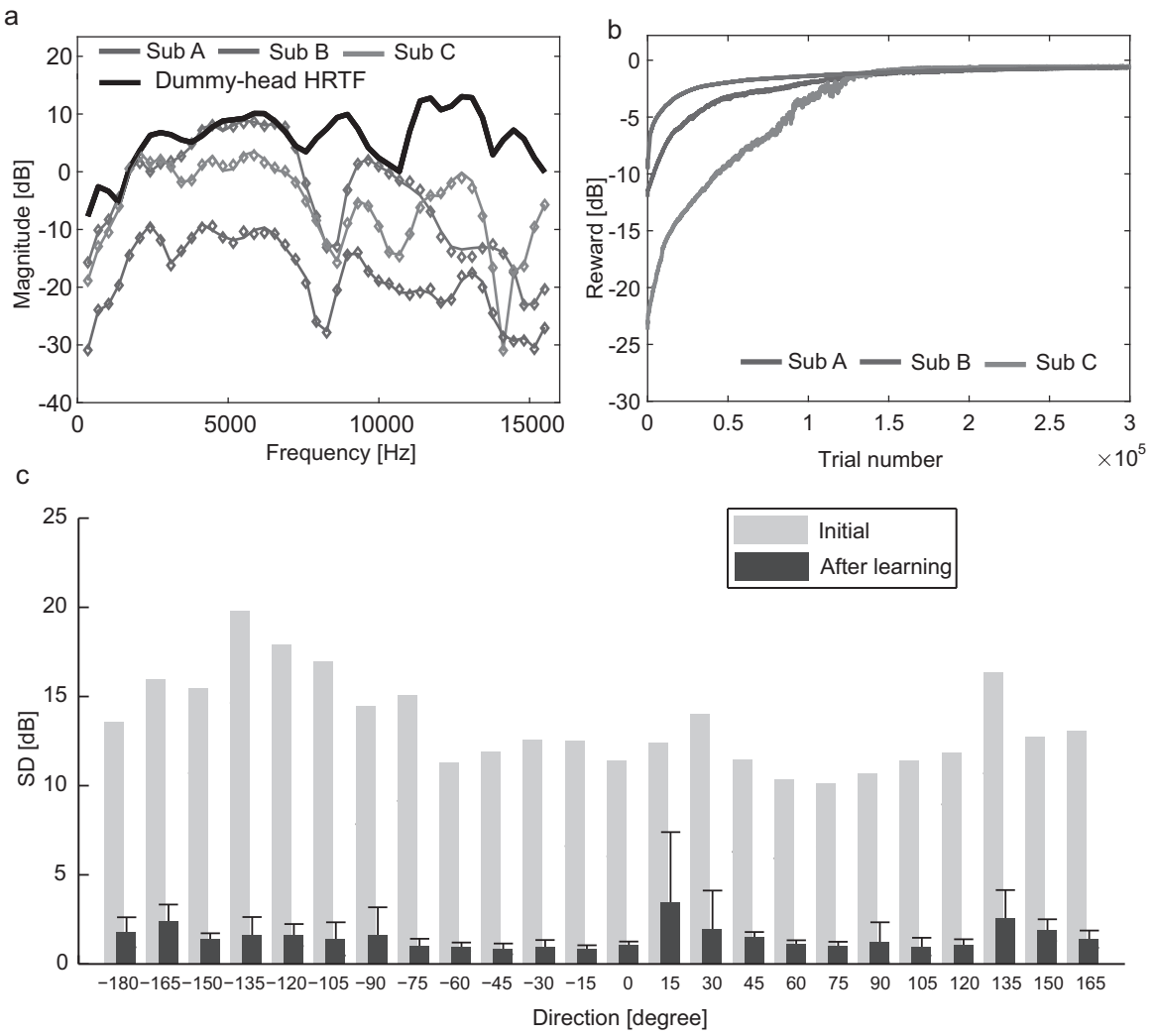


Fig. 2. Simulation results. (a) Learned HRTFs of the right ear in a 30° direction to the right for three representative subjects (Sub A, B, and C). A bold black line indicates the dummy-head HRTF (initial values), and colored lines respectively represent the learned HRTF for each subject (values after learning). The diamond markers denote the measured individual HRTFs (the targets). Each color corresponds to each subject. (b) Rewards along with the trials. The same color in the right panel is assigned for each subject. For illustration purposes, the reward was averaged and plotted for every 50 trials. (c) Initial and learned SDs for each direction in the horizontal plane. The SD shown is the average across subjects, and the error bar indicates its standard deviation for the subjects.

direction of the evaluation sound was the same as that of the reference. On the other hand, the dummy-head HRTF was poorly perceived because it was generated from a different direction. The average localization accuracy of the learned HRTF across six subjects was significantly larger than the one obtained from the dummy-head HRTF ($p < 0.01$). No significant difference was found between the localization accuracies of the measured and learned HRTF, which further indicates that both HRTFs were similar. Although the rating in Evaluation 2 for the learned HRTF was smaller, compared to that of the measured one ($p < 0.01$) (Fig. 3), this rating for the learned HRTF was still significantly better than the one for the dummy-head HRTF ($p < 0.05$). Moreover, its average rate was four, which indicates that the location of the sound generated from the learned HRTF was within 15° of the left/right field. Thus, these results suggest that the learned HRTF was sufficient to achieve better localization performance.

3 LEARNING EXPERIMENT

Next we performed a proof-of-concept learning experiment where an individual HRTF was acquired using our proposed method without a measured HRTF.

3.1 Protocol

Six subjects (ages 21 to 24) participated in this experiment, which was approved by the Ethics Committee of Nagaoka University of Technology. All subjects were given instructions about the study and provided informed consent before the experiment. The experiment was done in a sound-proof room where HRTF was measured. For each subject, we set one target direction (Table 1). The target direction was informed to the subject before the experiment. This was different for each subject and selected randomly from 24 directions used in the simulation. To simplify the experiment, we personalized (learned) one of the left and

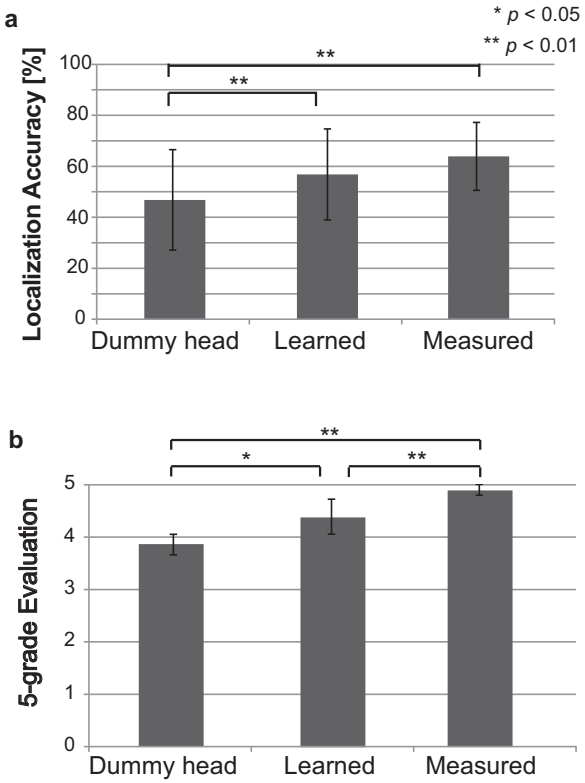


Fig. 3. Averaged localization accuracy (a) and evaluation grade (b) across all subjects. The error bar indicates the standard deviation for the different subjects.

Table 1. Evaluations after the learning experiment

	Azimuth direction (in degrees)				
Sub	Target	Learned (error)	Learned (error)	ΔDir	Score
1	30	25.5 (–4.5)	25.0 (–5.0)	0.5	0.65
2	–45	–39.5 (+5.5)	–33.5 (+11.5)	6.0	2.00
3	–45	–91.5 (–46.5)	–100.0 (–55.0)	8.5	1.05
4	–45	–61.0 (–16.0)	–48.5 (–3.5)	–12.5	–0.10
5	60	60.0 (0)	86.7 (+26.7)	26.7	1.10
6	–30	–28.5 (+1.5)	30.0 (+60)	58.5	1.05
Ave		(12.3) *	(26.7) *	14.3	0.96

* Averaged absolute directional errors across subjects.

right HRTFs from the dummy-head HRTF for the target direction while the other HRTF was fixed at the dummy-head HRTF. This simple setting allowed the subjects to perform the experiment without excessive fatigue.

The subjects wore intra-concha earphones (MDR-ED238, Sony, Japan) and the sound was amplified through a USB audio interface (UA-55, Roland, Japan). The experiment consisted of 300 trials. In a trial, a series of two virtual sounds (1000 ms in duration) were presented to the subject by convolving the HRIR with white noise (100 Hz–15 kHz) (Fig. 4(a)). First, the sound used in the previous trial was provided as a reference sound. At the same time, the evaluated direction (position) of the previous trial was also presented on the display. Then, the evaluation sound was generated from the learned HRTF and presented 500 ms

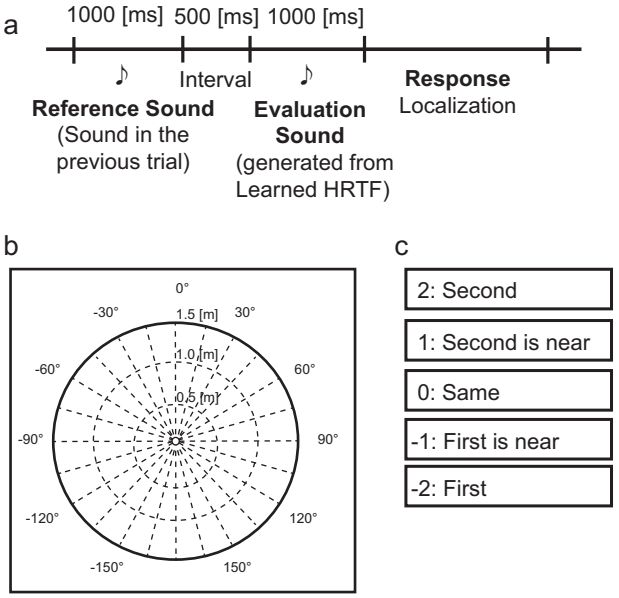


Fig. 4. (a) Experimental protocol for the learning experiment. Note that the sound was presented for 1000 ms with a 500-ms interval. (b) Display for localization responses in the azimuth direction evaluation. (c) Five criteria for the closeness evaluation. The subject was asked to answer which sound (first or second) is closer to the target.

after the first sound. In the first trial, the subjects were presented with only the evaluation sound, which was generated using the dummy-head HRTF. During the second trial, both the reference and evaluation sounds were presented.

After listening to the two sounds, the subjects evaluated the direction of the second sound (d_{answer}) in the horizontal plane by pointing to it using a graphical user interface (Fig. 4(b)). Using this subjective evaluation of the direction, the method updated the learned HRTF.

The reward for reinforcement learning was defined as:

$$r = -\frac{|d_{target} - d_{answer}|}{10}, \tag{4}$$

where d_{target} are the directions (in degrees) of the targets. This reward indicates directional difference between target and perceived sounds and was set so that r becomes –1 when the directional difference is 10°. Note that we modified an HRTFs ipsilateral to the target direction; the HRTF for the right ear was modified (learned) when the target was on the right side (positive angle), whereas that for the left ear was modified when the target was on the left side (negative angle).

3.2 Learning Results

The reward r obtained via the learning experiment exhibited a gradual decrease in magnitude for many subjects as the sound generated from the learned HRTF approached the target direction. The averaged r across the subjects in the first trial was –4.4 and decreased to –1.1 at the end of the trials. The learning experiment took approximately 30 minutes.

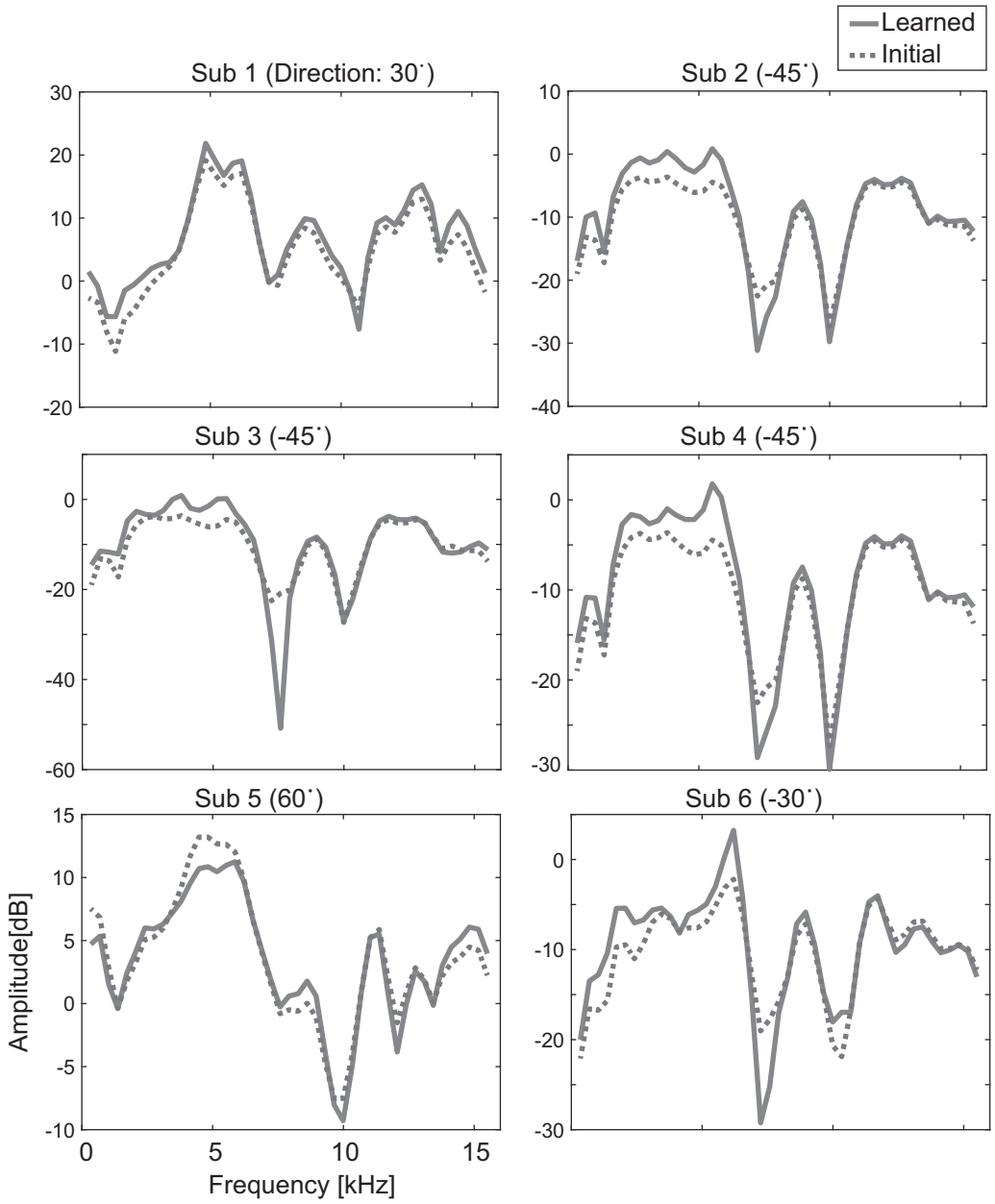


Fig. 5. Spectral difference between initial (dashed red line) and learned (blue bold line) HRTFs. Each panel shows the result for each subject. The direction of the target is shown in the parenthesis.

3.3 Post Evaluations (The Direction and the Closeness)

On a different day after conducting the experiment, the directions and proximity to targets for the sounds generated by the learned HRTF were assessed (post evaluations). In the direction evaluation, two types of virtual sounds, those generated using the dummy-head HRTF and the learned HRTF (after 300 trials), were presented randomly and were each evaluated 10 times using the interface (Fig. 4(b)). Improvement of the localization was defined as the directional difference between the learned and initial sound (ΔDir) in the horizontal plane. A positive ΔDir shows that the learned sound was closer to the target than the initial sound.

Next, in order to subjectively evaluate the sound generated from the learned HRTF, the closeness (proximity to targets) evaluation was conducted. The subjects were presented with the two sounds (the dummy-head and the learned sounds) consecutively. The sound presentation order was counter-balanced: The dummy-head sound was presented first (first presented sound) in the first half of the trials and the learned sound was presented first in the rest of trials. The subjects assessed the extent to which the sound was close to the target direction (with five graded values: –2, –1, 0, 1, and 2; Fig. 3(c)). A grade of –2 was awarded when only the first presented sound was heard from the target direction, while grade 2 was assigned when only the second sound was heard from the target direction. Grades 1 and –1 mean that the second sound was closer to the target than the first one, and vice versa, respectively. When the two sounds were heard from the same direction, a value of

0 was awarded. The evaluation was conducted 20 times in total. The subject was not informed which sound was generated using the learned HRTF. We converted the graded values into corresponding subjective values (referred to as the score), which represented the extent to which the sound generated using the learned HRTF was localized to the target. This was achieved by multiplying the graded values by –1 for the data when the first presented sound was generated using the dummy-head HRTF because meaning of the graded values was different depending on the first presented sound. We averaged this value across the trials for each subject and conducted statistical analysis using one-tailed Wilcoxon signed rank test ($n = 6$).

In both evaluations, we also presented sounds generated using the measured or learned HRTF after 150 trials. However, the results obtained were not used for the analysis based on the purpose of the experiment.

3.4 Post Evaluation Results

The results for the post evaluations revealed that the sound generated from the learned HRTF was closer to the target than that from the dummy-head HRTF for five out of six subjects (Table 1), although it still deviated from the target (by 12.3° on average). Similarly, the difference between learned and initial HRTFs (ΔDir) was positive (14.3°) on average. The scores obtained from the closeness evaluation also supported improvement in localization accuracy. A significantly positive score was obtained ($p < 0.05$, signed-rank: 20), suggesting that the sound generated from the learned HRTF was closer to the target.

In the result for subject 6, the initial direction of the sound was 30° while the target was in the opposite lateral direction (–30°). This is very difficult to interpret because of the left-right confusion that does not occur usually. However, even after excluding this subject, we observed a similar tendency; the average directional error was 14.5, ΔDir was 5.4, and the score was 0.94.

3.5 HRTF Spectral Changes

On comparing the spectral magnitude of learned HRTFs to the initial one (dummy-head), we found that the changes in the amplitude were observed at the frequencies of 2–8 kHz (Fig. 4). A decrease in the amplitude at frequencies of 7–8 kHz was observed for the data with a negative target angle (subjects 2, 3, 4, and 6), whereas this change was not observed for the data with a positive target angle (subjects 1 and 5). Also, amplitudes in the frequency range of 2–6 kHz are higher for data with a negative target but it was lower for subject 5 (target angle was 60°). Thus, HRTFs were differently modified depending on the target angle of the subject.

In this study we focused on the HRTF spectrum and removed the initial delay before the modeling (see Sec. 1.1). Therefore, we expected that the interaural time difference (ITD) was unchanged. When we examined the changes of the ITD after the learning, we found a small increase (just one sample; 0.023 ms) of ITD for two participants (subject 1 and 3). However, for the other participants we did not. This change of ITD corresponds to the 1.9° of ITDs calculated by theoretical ITD models (Kuhn 1977). Thus, changes of ITD resulting from the modification of the spectral phase of HRTF were not obvious from our results.

4 DISCUSSION

We proposed a new reinforcement learning method based on HRTF personalization to improve virtual sound localization for human listeners. To demonstrate its feasibility, we conducted a simulation and an experiment with human subjects (learning experiments). In simulations, we showed that our method can learn an individually personalized HRTF from a dummy-head HRTF. These results demonstrate that our method can be used for any direction and any user if the learning is completed. Furthermore, we confirmed the improvement of the localization accuracy for the learned HRTF in the learning experiment, where no measured HRTF for individuals was obtained. This result suggests that our method has potential for a practical application. We now discuss the advantages and problems of these results.

4.1 Advantage of Reinforcement-Learning-Based Personalization

The advantage of the proposed method is that any HRTF can be estimated based on the user's evaluation. When using another individual's HRTF that seems to be very similar in objective [4–7] or perceptual standpoint [12], an HRTF might be selected in which sound is not well localized owing to the differences of individual HRTFs. Even when combining or scaling other user's HRTFs, only a few parameters are searched [11,13]. In contrast, as described in the Introduction section, our method can estimate any HRTF from the subjective evaluation alone without any restriction. In fact, we found the changes of HRTF after the learning, which were different among the subjects. The average directional error compared to the target direction was 12.3°. Previous studies reported that the directional error was around 20° using the estimated HRTF (21° for [13] in the horizontal plane and 20.7° for [5] in the median plane). Thus, our results indicate a possibility that our method can be used to obtain better individual HRTFs.

In this study we used an ARMA model to represent HRTFs (HRIR) because this method effectively models HRTFs [16]. However, our method can use any other modeling method (e.g., PCA) in order to reduce the number of parameters modified by reinforcement learning. In addition, the reinforcement learning process can deal with such modeling parameters and other information, including anthropometric features, ITD, and directional information simultaneously. Thus, our proposed method is flexible in terms of the parameters to be searched, although there is a trade-off between the number of parameters for reinforcement learning and personalization (learning) speed. Recent studies have shown that intraconic components of HRTFs are more critical for accurate virtual sound localization than ITD or lateral spectral components [22, 23]; therefore,

incorporating such information into our method could lead to promising results.

This flexibility advantage also enables us to apply the proposed method to both the estimation of an individual HRTF and to further improvement of the measured HRTF. Owing to the difficulty of HRTF measurement, a localization error for the virtual sounds is usually observed, even when the sound is generated from the measured HRTF. In our previous work [24] we showed in a simulation setting that measured HRTF with added noise can be recovered by our reinforcement learning method. Thus, our proposed method can likewise be used to reduce such errors.

4.2 Issues to Be Solved for Rapid Personalization

Our current method has several limitations. First, it is necessary to achieve rapid personalization by reducing the number of trials because the learning speed was slow (Fig. 2(b)) and the improvement did not seem to be complete (Fig. 2(c)) in the simulation, probably due to the large search spaces of the HRTF model. Additionally, the personalization of the HRTF in the learning experiment of the present study took approximately 30 minutes for a single direction. This is longer than the time taken in previous studies [11, 12], where customization of HRTFs for multiple directions were performed within 20 minutes using subjective evaluation or selection of HRTFs. Although rewards used in the simulations and the learning experiment (i.e., SD and user's evaluation) both represent how much the current HRTFs differs from the appropriate HRTFs, the user's evaluation might be more variable or noisy than SD because of the variability human perception in the participants. This could result in lower learning speed during the learning experiment.

One solution would be to select an appropriate initial HRTF instead of the dummy-head HRTF. Here we started HRTF personalization from the dummy-head HRTF. This was because the dummy-head represents an average-size human upper body and head. Therefore, the difference between the dummy-head HRTF and individual HRTF to be estimated may have been minimized on average. However, the difference varied among subjects. This affected the number of trials necessary for obtaining an individual HRTF in the simulation. The larger the initial SD, the longer the time necessary for success. Additionally, the SD did not become zero in the simulation, and hence, there exists a small error or difference as compared to the measured HRTFs (Fig. 2(c)). We speculate that this happens because the dummy-head HRTF deviates from individual HRTFs and perhaps parameters of ARMA modeling are specific to the dummy-head HRTF. Therefore, it was expected that a different initial HRTF promoted quick learning of the individual HRTF. We expect that rapid personalization can be achieved by selecting an initial HRTF using the above-mentioned previous approaches, which select an HRTF based on the similarity of anthropometric features [4–8] or generate a simulated HRTF by numerical analysis techniques based on morphological information [9,10].

There are some limitations to our experimental setting. Our proposed method focused on the modification of the HRTF spectrum and it obtained improvements in localization accuracy (Table 1), suggesting the effectiveness of our method and importance of the HRTF spectrum in the horizontal-plane localization. However, it has been well-known that in addition to the spectral changes of HRTFs, ITD, and interaural level differences (ILD) are important for accurate sound localization in the horizontal plane [1]. In addition, the degrading accuracy due to the non-individual HRTFs are likely to be relatively small [22]. Therefore, improvements of the sound localization could be limited to some extent in our current settings. Considering the rest of cues (ITD and ILD), as well as the HRTF spectrum, may promote efficient and rapid personalization. We must also examine the feasibility of our method in a more practical scenario by including the median plane or the sagittal plane. It is necessary to adopt a setting enabling personalization for multiple directions simultaneously. Because the current experimental setup was very simple, and the HRTF was modified for one ear and one direction only.

The adaptation to a specific sound could be another factor to improve the localization performance. A previous study has shown that the human listener improved the localization accuracy for sound generated by non-individual HRTFs [25]. Effects of the adaptation in our results could be small because we evaluated the improvements of the localization on a different day of the learning experiment (see Sec. 3.3). However, one of the subjects' performance degraded post evaluation (Table 1), even though the reward was decreased during learning. This could be because this subject adapted to the sound during learning and the adaption during learning affected user's evaluation.

Thus, our method has the possibility of obtaining personalized HRTFs; however, it currently may not work well for everyone. To solve this issue, it may be necessary to examine the details of the learned HRTFs, i.e., SDs and other measures reflecting human auditory systems [20, 21] as well as effects of the adaptation on user's evaluation. These topics should be addressed in future work.

5 CONCLUSION

In this paper we proposed a method based on reinforcement learning to obtain a personalized HRTF and tested the feasibility of the method. The results may serve as a proof-of-concept of our method. After improving learning speed and testing in more practical settings, our method is expected to be applied to easily obtain a personalized HRTF at any location.

6 ACKNOWLEDGMENTS

This work was supported by Nagaoka University of Technology Presidential Research Grant and JSPS KAKENHI Grant Numbers 24300051, 24650104, and 16K00182. We would like to thank Editage (www.editage.jp) for English language editing.

7 REFERENCES

[1] H. Möller, "Fundamentals of Binaural Technology," *Appl. Acoust.*, vol. 36, pp. 171–218 (1992).

[2] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization Using Nonindividualized Head-Related Transfer Functions," *J. Acoust. Soc. Am.*, vol. 94, pp. 111–123 (1993 July). https://doi.org/10.1121/1.407089.

[3] S. Yano, H. Hokari, and S. Shimada, "A Study on Personal Difference in the Transfer Functions of Sound Localization Using Stereo Earphones," *IEICE Transactions on Fundamentals of Electronics Communications and Computer Sciences E Series A*, vol. 83, pp. 877–87 (2000).

[4] D. Zotkin, J. Hwang, R. Duraiswaini, and L. S. Davis, "HRTF Personalization Using Anthropometric Measurements," IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 157–160 (2003). https://doi.org/10.1109/ASPAA.2003.1285855.

[5] K. Iida, Y. Ishii, and S. Nishioka, "Personalization of Head-Related Transfer Functions in the Median Plane Based on the Anthropometry of the Listener's Pinnaea)," *J. Acoust. Soc. Am.*, vol. 136, pp. 317–333 (2014 July). https://doi.org/10.1121/1.4880856.

[6] S. Xu, Z. Li, and G. Salvendy, "Improved Method to Individualize Head-Related Transfer Function Using Anthropometric Measurements," *Acoust. Sci. & Tech.*, vol. 29, pp. 388–390 (2008). https://doi.org/10.1250/ast.29.388.

[7] T. Nishino, N. Inoue, and K. Takeda, "Estimation of HRTFs on the Horizontal Plane Using Physical Features," *Appl. Acoust.*, vol. 68, pp. 897–908 (2007). https://doi.org/10.1016/j.apacoust.2006.12.010.

[8] E. A. Torres-Gallegos, F. Orduña-Bustamante, and F. Arámbula-Cosío, "Personalization of Head-Related Transfer Functions (HRTF) Based on Automatic Photo-Anthropometry and Inference from a Database," *Appl. Acoust.*, vol. 97, pp. 84–95 (2015 Oct.). https://doi.org/10.1016/j.apacoust.2015.04.009.

[9] C. T. Jin, P. Guillon, N. Epain, R. Zolfaghari, A. van Schaik, A. I. Tew, C. Hetherington, and J. Thorpe, "Creating the Sydney York Morphological and Acoustic Recordings of Ears Database," *IEEE Trans. Multimedia*, vol. 16, pp. 37–46 (2013 Dec.). https://doi.org/10.1109/TMM.2013.2282134.

[10] P. Mokhtari, H. Takemoto, R. Nishimura, H Kato, "Computer Simulation of KEMAR's Head-Related Transfer Functions: Verification with Measurements and Acoustic Effects of Modifying Head Shape and Pinna Concavity," *Principles and Applications of Spatial Hearing* (World Scientific, 2011), pp. 205–215.

[11] J. C. Middlebrooks, E. A. Macpherson, and Z. A. Onsan, "Psychophysical Customization of Directional Transfer Functions for Virtual Sound Localization," *J. Acoust. Soc. Am.*, vol. 108, pp. 3088–3091 (2000 Dec.). https://doi.org/10.1121/1.1322026.

[12] Y. Iwaya, "Individualization of Head-Related Transfer Functions with Tournament-Style Listening Test: Listening with Other's Ears," *Acoust. Sci. & Tech.*, vol. 27, pp. 340–343 (2006). https://doi.org/10.1250/ast.27.340.

[13] K. J. Fink and L. Ray, "Individualization of Head Related Transfer Functions Using Principal Component Analysis," *Appl. Acoust.*, vol. 87, pp. 162–173 (2015 Jan.). https://doi.org/10.1016/j.apacoust.2014.07.005.

[14] K. Doya, "Reinforcement Learning in Continuous Time and Space," *Neural Computation*, vol. 12, pp. 219–45 (2000).

[15] J. Morimoto and K. Doya, "Acquisition of Stand-Up Behavior by a Real Robot Using Hierarchical Reinforcement Learning," *Robotics and Autonomous Systems*, vol. 36, pp. 37–51 (2001). https://doi.org/10.1016/S0921-8890(01)00113-0.

[16] Y. Haneda, S. Makino, Y. Kaneda, and N. Kitawaki, "Common-Acoustical-Pole and Zero Modeling of Head-Related Transfer Functions," *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 188–196 (1999). https://doi:10.1109/89.748123.

[17] K. Watanabe, R. Kodama, S. Sato, S. Takane, and K. Abe, "Influence of Flattening Contralateral Head-Related Transfer Functions upon Sound Localization Performance," *Acoust. Sci. & Tech.*, vol. 32, pp. 121–124 (2011). https://doi:10.1250/ast.32.121.

[18] D. Morikawa and T. Hirahara, "Signal Frequency Range Necessary for Horizontal Sound Localization," *Acoust. Sci. & Tech.*, vol. 31, pp. 417–419 (2010). https://doi:10.1250/ast.31.417.

[19] K. Matsui and A. Ando, "Estimation of Individualized Head-Related Transfer Function Based on Principal Component Analysis," *Acoust. Sci. & Tech.*, vol. 30, pp. 338–347 (2009). https://doi:10.1250/ast.30.338.

[20] R. Baumgartner, P. Majdak, and B. Laback. "Modeling Sound-Source Localization in Sagittal Planes for Human Listeners," *J. Acoust. Soc. Am*, vol. 136, pp. 791–802 (2014 Aug.).

[21] R. Baumgartner, and P. Majdak. "Modeling Localization of Amplitude-Panned Virtual Sources in Sagittal Planes," *J. Audio Eng. Soc.*, vol. 63, pp. 562–569 (2015 Jul./Aug.).

[22] G. D. Romigh and B. D. Simpson, "Do You Hear Where I Hear?: Isolating the Individualized Sound Localization Cues," *Front Neurosci*, vol. 8, p. 370 (2014). https://doi:10.3389/fnins.2014.00370.

[23] F. Grijalva, L. Martini, D. Florencio, and S. Goldenstein, "A Manifold Learning Approach for Personalizing HRTFs from Anthropometric Features," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, pp. 559–570 (2016 Feb.). https://doi:10.1109/TASLP.2016.2517565.

[24] S. Morioka, I. Nambu, S. Yano, H. Hokari, and Y. Wada, "Adaptive Modeling of HRTFs Based on Reinforcement Learning," *Lecture Notes in Computer Science (ICONIP2012)*, vol. 7666, pp. 423–430 (2012 Feb.).

[25] F. Klein and S. Werner, "Auditory Adaptation to Non-Individual HRTF Cues in Binaural Audio Reproduction," *J. Audio Eng. Soc.*, vol. 64, pp. 45–54 (2016 Jan./Feb.), https://doi.org/10.17743/jaes.2015.0092.

8 APPENDIX 1: DETAILS OF REINFORCEMENT LEARNING ALGORITHM

We use the actor-critic reinforcement learning method, which is a temporal-difference (TD) learning method. The actor selects the appropriate coefficients of ARMA models as outputs. The critic evaluates the values of the selected coefficients (i.e., states) and calculates the TD error δ_{TD}. This error is used for updating the actor and critic functions.

In this study we used continuous state modeling [14, 15] of the coefficients for the discrete time, which enabled fine-tuning of the HRTF. For further details of the algorithm and parameters, see [14, 15].

8.1 Critic

Using the current state of the ARMA coefficients, x[t], the critic evaluated the value function at trial iteration t, $V(x[t])$, as follows:

$$x\,[t] = \left[p\,(1)\,,\ p\,(2)\,,\cdots,\,p\,(P)\,,q\,(0)\,,q\,(1)\,,\right.$$
$$\left.\cdots,\,q\,(Q)\right],\ V\ (x\,[t]) = \sum_i v_i b_i\,(x\,[t])\ , \quad (5)$$

where v_i is a weight and $b_i()$ is a basis function. The parameter p and q are AR and MA parameters, and P and Q are the number of AR and MA parameters, respectively. The value function $V(x[t])$ is a linear combination of the weight v_i and basis function b_i. Based on $V(x[t])$, the TD error is

$$\delta_{TD}\,[t]\ = r\,[t] + \gamma V\,(x\,[t]) - V\,(x\,[t-1])\,. \quad (6)$$

Here, $r[t]$ is the reward for the reinforcement learning at trail t, and the reward (in the simulation) is defined as a spectrum distortion. $\gamma = 0.95$ is a discount rate. Using this TD error, the weight is updated as

$$v_i = \alpha\dot{\delta}_{TD}e_i\,(t)\,, \quad (7)$$

where α is the learning rate, e_i is an eligibility trace, and $\dot{\delta}_{TD}[t] = \delta_{TD}\,[t] - \delta_{TD}[t-1]$. To avoid learning by the actor when the critic cannot properly evaluate the values of the current state, α is set depending on the trial number t as

$$\alpha\ = min\left[\alpha_c,\ \frac{\alpha_c}{2}\left(1+\frac{t}{T}\right)\right]. \quad (8)$$

Here $\alpha_c = 0.12$ is a predefined learning rate and $T = 1{,}000$ is the number of trials to use α_c. The eligibility trace is updated by the following rule:

$$\dot{e}_i\,[t+1] = \gamma\lambda e_i\,[t] + b_i\,(x\,[t])\,, \quad (9)$$

where $\lambda = 0.91$ is the decay parameter of the trace.

8.2 Actors

The actor is used to update the ARMA modeled parameters as follows:

$$u_j\,[t]\ = {u_j}^{max}\ g\left(\sum_i w_{ij}b_i\,(x\,[t]) + \sigma n_j\,[t]\right)$$
$$+\,{u_j}^{bias}, \quad (10)$$

where $u_j[t]$ is the j-th component of the updated ARMA parameters $u[t]$, ${u_j}^{max} = 2$ denotes the maximum values of u_j, and ${u_j}^{bias} = -1$ is a bias (constant). In the present study, outputs of the actor $u[t]$ become the states in the next trial ($x[t+1]$). In addition, w_{ij} is a weight and $g(z) = 1/1 + e^{-sz}$ is a sigmoid function with gain s ($= 2.2$). $n_j[t]$ is the Gaussian noise for the j-th component of $u[t]$ generated from the Gaussian distribution $n_0 \times \mathcal{N}(0, 1)$. The search noise is defined as $\sigma n_j[t]$ and the magnitude of the search noise σ is set as

$$\sigma\ = \sigma_0\,min\left[1,\ max\left[0,\ \frac{V_1 - V\,(x\,[t])}{V_1 - V_0}\right]\right], \quad (11)$$

where $\sigma_0 = 1$ is the noise gain and $V_1 = 0$ and $V_o = -100$ are the maximum and minimum of $V(x[t])$, respectively. The update rule for the weight is

$$w_{ij} \leftarrow w_{ij} + \beta\delta_{TD}\,[t]\,\sigma n_j\,[t]\,b_i\,(x\,[t])\,. \quad (12)$$

The weight learning rate for the actor is modulated depending on t as

$$\beta\ = min\left[\beta_c,\ \frac{\beta_c}{2}\left(1+\frac{t}{T}\right)\right]. \quad (13)$$

Both α_c and β_c are set to the same value (0.12). The basis function $b_i()$ is modeled by the incremental normalized Gaussian network (INGnet) [3] as

$$b_i\,(x\,[t])\ = \frac{a_i\,(x\,[t])}{\sum_{l=1}^{K} a_l\,(x\,[t])}\ , \quad (14)$$

where $a_i(x[t])$ is called the activation function and is defined by

$$a_i\ (x\,[t]) = \exp\left(-\frac{1}{2}M_k(x\,[t]-c_k)^2\right). \quad (15)$$

Here, *exp* is the exponential function, M_k is the full width at half maximum set as 2.5 for all k, and c_k is the center of the Gaussian function. Using INGnet, the model adds a new intermediate node (for the critic and the actor) according to the amount of TD errors. A new node in the intermediate layer for the critic is added when the following condition is met:

$$\delta_{TD}\,[t] > e_{max}\ and\ \ max_i a_i\,(x\,[t]) < \ a_{min} \quad (16)$$

Similarly, a new node is added for the actor under the following condition:

$$max\left|\delta_{TD}\,[t]\,\sigma n_j\,[t]\right| > e_{max}\ and\ max_i a_i\,(x\,[t])$$
$$<\ a_{min}\ . \quad (17)$$

In this case, the intermediate layer is initialized as $c_k = x[t]$ and $w_{kj} = x_j\ [t] + \delta_{TD}[t]\sigma n_j[t]$. a_{min} is set to 0.4. To avoid unnecessary incrementing of the node during the initial phase of the learning on account of the incomplete weight learning, the threshold e_{max} is defined as

$$e_{max} = \ 0.4\ exp\left(-\frac{t}{T}\right). \quad (18)$$

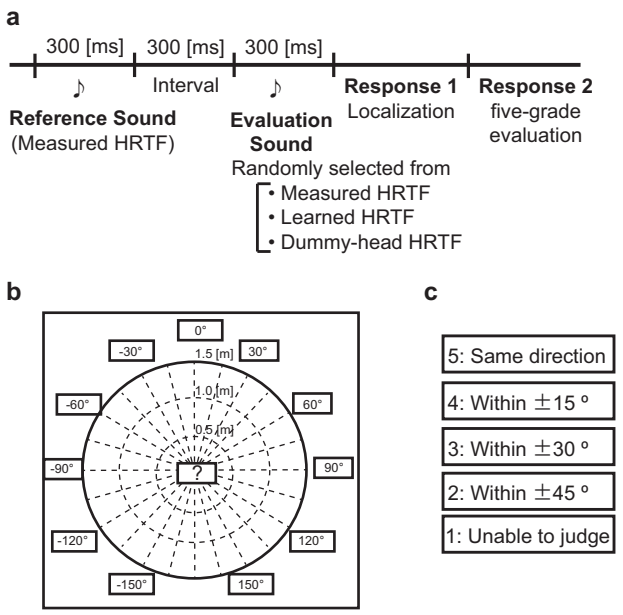


Fig. 6. (a) Experimental protocol for the evaluation experiment. (b) Display for localization responses in Evaluation 1. The question mark located at the center of the display was to be pressed when the subject was unable to identify the direction. (c) Five-grade evaluation criteria for comparison with the reference in Evaluation 2. These criteria were shown in the display and the subject was asked to click the button (number) corresponding to the evaluation.

9 APPENDIX 2: DETAILS OF EVALUATION EXPERIMENT FOR LEARNED HRTF

We performed an experiment with human subjects to show that the learned HRTF in the simulation was actually used for sound localization. This study was approved by the Ethics Committee of Nagaoka University of Technology. All the human subjects were given instructions concerning the study and their informed consent was taken before the experiment. The experiment was conducted in a soundproof room where the HRTF was measured.

9.1 Protocol Details

Similar to the learning experiment, the subjects wore intra-concha earphones and the sound was amplified through a USB audio interface. A series of two virtual sounds were presented to the subjects (Fig. 6(a)). The first sound served as a reference; the second was an evaluation sound. The sound was prepared by convoluting HRIR with white noise, and its direction was selected from 12 directions in 30° intervals. The subject was instructed to listen to these two sounds and evaluate whether the second sound was generated from the same direction as the first one. The measured HRTF for each subject was used as the reference sound. For the evaluation sound, either the learned HRTF from the simulation or the dummy-head HRTF was used. The HRTF with the lowest SD at the end of learning in the simulation was selected and used as the learned HRTF. The order in which these two sounds were presented was pseudo-randomized. A trial consisted of hearing two sounds (the reference and evaluation sounds) and two subjective evaluations (sound localization in Evaluation 1 and five-grade evaluations in Evaluation 2). A single session consisted of 72 trials; therefore, 5 sessions were conducted for a total of 360 trials. Measurements of 10 sets of evaluations were conducted for each sound direction.

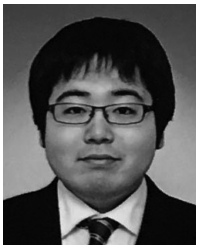
In Evaluation 1, the subjects were asked to immediately determine the direction of the sound (Fig. 6(b)). In the second evaluation (Evaluation 2: five-grade evaluation), the subjects were asked to assess the extent to which the direction of the second (evaluation) sound deviated from that of the first sound (reference) using five criteria (Fig. 6(c)). We expected that the five-grade evaluation would yield more detailed information about the comparisons of the two sequential sounds because the comparison was easier than reporting the direction of the sound. Using these two evaluations, we aimed to confirm sound localization in a real setting with the learned HRTF via reinforcement learning. To test the statistical significance of the localization accuracy (Evaluation 1) and five-grade evaluation (Evaluation 2), a paired t-test ($n = 6$) was performed for each pair of the evaluation sounds (measured, learned, and dummy-head HRTFs).

THE AUTHORS


Isao Nambu


Manabu Washizu


Shuhei Morioka


Yuta Hasegawa


Wataru Sakuma


Shohei Yano


Haruhide Hokari


Yasuhiro Wada

Isao Nambu received his doctoral degree in engineering from the Nara Institute of Science and Technology in 2010. From 2012 to 2017 he was an assistant professor at the Nagaoka University of Technology. Since 2017, he is an associate professor at the Nagaoka University of Technology. His research interests include neuroimaging, brain-computer interfaces, and human auditory processing.

•

Manabu Washizu graduated from the Nagaoka National College of Technology in 2011. From 2011 to 2015 he was a member of the neural information processing laboratory at Nagaoka University of Technology. He received his bachelor and master degrees from Nagaoka University of Technology in 2013 and 2015, respectively. His research interest is the personalization of spatial audio environment.

•

Shuhei Morioka graduated from the Nagaoka University of Technology and received his master degree in engineering. From 2011 to 2013 he studied the personalization of head-related transfer function using reinforcement learning. Since 2013 he works as a web engineer.

•

Yuta Hasegawa graduated from the Nagaoka National College of Technology in 2013 and received his bachelor and master degrees in engineering from the Nagaoka University of Technology in 2014 and 2016, respectively. His research interest is the personalization of head-related transfer function using principal component analysis and particle swarm optimization.

•

Wataru Sakuma graduated from National Institute of Technology, Nagano College in 2014 and received his bachelor and master degrees in engineering from the Nagaoka University of Technology in 2016 and 2018, respectively. His research interest is perception information processing.

•

Shohei Yano received his doctoral degree in engineering from the Nagaoka University of Technology in 2000. From 2000 to 2010 he was an assistant professor with National Institute of Technology, Nagaoka College. Since 2015 he has been an associate professor at National Institute of Technology, Nagaoka College. In 2016 he was an associate professor at Nagaoka University of Technology. His research mainly focuses on the audio signal processing. He is a member of the Institute of Electronics, Information and Communication Engineers Japan and the Acoustical Society of Japan.

•

Haruhide Hokari was born in Nagaoka, Japan, in 1949. He graduated from the Mechanical Engineering Department of Nagaoka Technical Senior School, Nagaoka, Japan, in 1969. He was a technical official of the Niigata University until 1979. He joined the Nagaoka University of Technology in 1979, where he is a technical official. From 2008 to 2010 he was a technical manager. His current research includes HRTF, stereophonic reproduction, and psychoacoustics.

•

Yasuhiro Wada received his B.E. and M.E. degrees in engineering from Tokyo Institute of Technology in 1980 and 1982 and then joined Kawasaki Steel Co. He was temporarily transferred to ATR. He obtained a D.Eng. degree in 1994. He then moved to the Nagaoka University of Technology, where he is currently a professor. His research interests include neural network models, motor learning control, and brain computer interface. He is a member of the Society for Neuroscience (SFN), the IEEE Engineering in Medicine and Biology Society (IEEE EMBS), the Japanese Neural Network Society (JNNS), the Institute of Electronics, Information and Communication Engineers (IEICE), and the Society of Instrument and Control Engineers (SICE).